UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

February 28, 2001

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES B-19*

MEMORANDUM FOR Howard Hogan
Chief, Decennial Statistical Studies Division

From: Donna Kostanich DK
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Mary H. Mulry
Bruce D. Spencer
ABT Associates Inc.
Contract No. 50-YABC-7-66020
Task No.46-YABC- 0-00004

Subject: Overview of Total Error Modeling and Loss Function Analysis

The attached document was prepared, per your request, to assist the Executive Steering Committee on A.C.E. Policy in assessing the data with and without statistical correction.

This report focuses on the methodology and assumptions used to assess total error and to compare uncorrected Census 2000 counts with corrected estimates based on the Accuracy and Coverage Evaluation Survey.

Overview of Total Error Modeling and Loss Function Analysis

Mary H. Mulry and Bruce D. Spencer

February 28, 2001

Contents

1. Improvement of Accuracy 4

2. Overview of Adjusted Estimates 6

3. Sources of Error in Adjustments 7

4. Total Error Modeling 14

References 17

Appendix A: Use of 1990 Data to Estimate Data Bias in the A. C. E. 19

Appendix B: Estimation of Data Bias 25

Appendix C: Imputation Error 36

Appendix D: Correlation Bias 37

Appendix E: Alternative Loss Functions 41

Appendix F: Loss Function Calculation 45

Appendix G: Inconsistency of Poststratification between P Sample and E Sample 51

Appendix H: Confidence Intervals 55

Tables

1. Evaluation Poststrata 57

2. Data Sources for Component Errors 58

3. Component Errors for Evaluation Poststratum 1 59

4. Component Errors for Evaluation Poststratum 2 60

5. Component Errors for Evaluation Poststratum 3 61

6. Component Errors for Evaluation Poststratum 4 62

7. Component Errors for Evaluation Poststratum 5 63

8. Component Errors for Evaluation Poststratum 6 64

9. Component Errors for Evaluation Poststratum 7 65

10. Component Errors for Evaluation Poststratum 8 66

11. Component Errors for Evaluation Poststratum 9 67

12. Component Errors for Evaluation Poststratum 10 68

13. Component Errors for Evaluation Poststratum 11 69

14. Component Errors for Evaluation Poststratum 12 70

15. Component Errors for Evaluation Poststratum 13 71

16. Component Errors for Evaluation Poststratum 14 72

17. Component Errors for Evaluation Poststratum 15 73

18. Component Errors for Evaluation Poststratum 16 74

19. 1990 & 2000 Undercount Rates 75

20. Total Error of Net Undercount Rate Assuming No Correlation Bias 76

21. Total Error of Net Undercount Rate Assuming No Correlation Bias for Nonblack Males 77

22 Total Error of Net Undercount Rate Assuming No Correlation Bias for 18-29 Nonblack Males .. 78

23 Total Error of Net Undercount Rate Including Correlation Bias of 2% Overcount for 18-29 Nonblack Males 79

24. Individual Effect of Errors on Bias, Standard Deviation, and Root Mean Square Error of Undercount Rate for the U.s. When All Other Errors Are Assumed to Be Zero 80

Figures

1. 95% Confidence Interval for Net Undercount Rate Assuming No Correlation Bias

2. 95% Confidence Interval for Net Undercount Rate Assuming No Correlation Bias for Nonblack Males

3. 95% Confidence Interval for Net Undercount Rate Assuming No Correlation Bias for 18-29 Nonblack Males

4. 95% Confidence Interval for Net Undercount Rate Including Correlation Bias of 2% Overcount for 18-29 Nonblack Males

5a. 1990 & 2000 Undercount Rates Corrected for Bias for Evaluation Poststrata and All Component Errors

5b. 1990 & 2000 Undercount Rates and Biases for Evaluation Poststrata and All Component Errors

1. Improvement of Accuracy

The census counts should be adjusted for undercount if the adjustment leads to congressional districts within each state that are more equal in numbers of persons. A variety of measures of inequality could be considered. The criterion we advocate essentially specifies that adjustment improves equality of district sizes when the sum of the mean squared errors (MSEs) for estimated district sizes is larger for the unadjusted estimates than for the adjusted estimates when considered across all states. This criterion has the properties that (i) an error in a Congressional district has approximately the same weight regardless the state to which it belongs, and (ii) a state's contribution to the overall measure of error is zero if all of the districts in the state are equal in actual size, regardless of error in the estimate of the state's population total. To apply this criterion, it is reasonable to use the existing districts (drawn after the 1990 census) and to define a measure of improvement from adjustment, say Δ, with

$$\Delta = n^{-1} \sum_{1 \le i \le 51} \tilde{P}_i^2 \sum_{1 \le j \le n_i} (MSE_{unadj,ij} - MSE_{adj,ij})$$

and n_i the number of congressional districts in state i (where for these purposes we include the District of Columbia as a state), $n = \sum_{1 \le i \le 51} n_i$ the total number of districts (or representatives), $\tilde{P}_i$ the size of state i, and $MSE_{unadj,ij}$ and $MSE_{adj,ij}$ the mean squared errors of the estimated proportion (or share) of state population i that is in district j based on unadjusted and on adjusted estimates, respectively. The measure Δ should not vary with minor changes in $\tilde{P}_i$ and to avoid undue complexity the unadjusted census count for state i may be used for $\tilde{P}_i$. Adjustment improves accuracy, and improves the equality of district sizes, if and only if Δ is greater than zero. (In the language of statistical decision theory, the measure Δ is equal to the difference in expected values of two loss functions.) For motivation behind the choice of Δ, see Spencer (2001). For further discussion of the interpretation of Δ, see Appendix E.

The statistical properties of the adjusted and unadjusted estimates must themselves be estimated,

and the process for doing so is called "total error modeling". The total error model is the basis for forming an estimate of Δ, say $\hat{\Delta}$, in what has been called a "loss function analysis". The development of $\hat{\Delta}$ depends on estimates of bias and variance, as discussed below. Once the estimate $\hat{\Delta}$ is available, unless the estimate is viewed as too untrustworthy the "loss function analysis" prescribes adjustment if $\hat{\Delta}$ is positive and non-adjustment if $\hat{\Delta}$ is negative. (This prescription eschews hypothesis testing, for reasons discussed in Spencer (2001).)

We now present a formal overview of the construction of $\hat{\Delta}$ based on results of total error modeling. Let E(.) and V(.) denote expectation and variance, respectively. Let U_{ij} denote the effect of undercount for population share for district ij (i.e., the fraction of state i population that is in district j) and let its estimate, the difference between the adjusted and unadjusted census count, be denoted by $\hat{U}_{ij}$. Denote the estimate of the variance $V_{ij} = V(\hat{U}_{ij})$ by $\hat{V}_{ij}$. Write the expected value of the undercount estimate as $E(\hat{U}_{ij}) = U_{ij} + B_{ij}$. Denote the estimate of B_{ij} by $\hat{B}_{ij}$. The measure of improvement of area ij from adjustment is $\Delta_{ij} = \tilde{P}_i^2(U_{ij}^2 - (B_{ij}^2 + V_{ij}))$, and the estimate of Δ_{ij} is taken to be $\hat{\Delta}_{ij} = \tilde{P}_i^2((\hat{U}_{ij} - \hat{B}_{ij})^2 - \hat{B}_{ij}^2 - 2\hat{V}_{ij})$. It may be shown that if $E(\hat{V}_{ij}) = V_{ij}$ and $E(\hat{B}_{ij}) = B_{ij}$ then $E(\hat{\Delta}_{ij}) = \Delta_{ij}$.

Total error analysis is used to analyze the propagation of the diverse sources of error in the adjusted estimates and to construct $\hat{B}_{ij}$; see Mulry and Spencer (1993, 1991) for details. It is reasonable to allow for the possibility that $\hat{B}_{ij}$ is biased, say $E(\hat{B}_{ij}) = B_{ij} - \beta_{ij}$, and it is reasonable to assume that the sampling designs are such that the correlation between $\hat{B}_{ij}$ and $\hat{U}_{ij}$ is negligible. The within-state average of Δ_{ij} is $\Delta_i = n_i^{-1} \sum_{1 \le j \le n_i} \Delta_{ij}$ and the weighted average across states is $\Delta = \sum_{1 \le i \le 51} (n_i/n)\Delta_i$. Similarly, the estimate for state i is $\hat{\Delta}_i = n_i^{-1} \sum_{1 \le j \le n_i} \hat{\Delta}_{ij}$ and the weighted average across states is $\hat{\Delta} = \sum_{1 \le i \le 51} (n_i/n)\hat{\Delta}_i$.

To partially summarize, notice that an estimate of improvement in accuracy, $\hat{\Delta}$, has been developed, so that adjustment appears to improve accuracy if and only if $\hat{\Delta} > 0$. To construct $\hat{\Delta}$ the Bureau requires estimates of undercount $\hat{U}_{ij}$, estimates $\hat{B}_{ij}$ of the bias of $\hat{U}_{ij}$, and estimates $\hat{V}_{ij}$ of the

variance of $\hat{U}_{ij}$. Biases β_{ij} in the estimates of bias $\hat{B}_{ij}$ may affect $\hat{\Delta}$. At the same time, the unmeasured bias affects the estimates of accuracy of the unadjusted estimates. To understand the net effect on the *comparative* accuracy, it is useful to observe (Spencer 2001) that the bias in $\hat{\Delta}_i$ is

$$E(\hat{\Delta}_i - \Delta_i) = 2n_i^{-1}\tilde{P}_i^2\sum_{j=1}^{n_i}\beta_{ij}E(\hat{U}_{ij}), \tag{4}$$

which is proportional to the cross-area correlation between β_{ij} and $E(\hat{U}_{ij})$ in state i, and that the bias overall is

$$E(\hat{\Delta} - \Delta) = 2n^{-1}\sum_{i=1}^{51}\tilde{P}_i^2\sum_{j=1}^{n_i}\beta_{ij}E(\hat{U}_{ij}). \tag{5}$$

This shows that the effect of omitting components of bias from the loss function analysis can be to favor adjustment or to favor non-adjustment, depending on the signs of the correlations between β_{ij} and $E(\hat{U}_{ij})$. See Appendix E for further discussion of loss functions.

2. Overview of Adjusted Estimates

Define the following notation for each poststratum, h.

$N_{C,h}$ = census "count" for poststratum h

$N_{C,h,ij}$ = census "count" for poststratum h in district j in state i

$I_{C,h}$ = number of persons imputed into the original enumeration for poststratum h

$\hat{I}_{E,h}$ = estimated number of enumerations in poststratum h with insufficient information for matching[1]

$\hat{E}_{E,h}$ = estimated number of erroneous enumerations in poststratum h

$\hat{N}_{CE,h}$ = estimated population size for poststratum h who could possible be matched

[1]Late enumerations are included with imputations in the original enumeration.

$\hat{N}_{CE,h} = N_{C,h} - I_{C,h} - \hat{I}_{E,h} - \hat{E}_{E,h}$

$\hat{N}_{P,h}$ = estimated size of the P sample population

$\hat{N}_{CP,h}$ = estimated number of the P sample population enumerated in the census

We estimate the population size N_h in poststratum h by

$$\hat{N}_h = \hat{N}_{CE,h}\hat{N}_{P,h}/\hat{N}_{CP,h}.$$

The adjustment factor for poststratum h is defined as $\hat{A}_h = \hat{N}_h/N_{Ch}$. The unadjusted estimate for district j in state i is $N_{unadj,ij} = \sum_h N_{C,h,ij}$ and the adjusted estimate is $\hat{N}_{adj,ij} = \sum_h \hat{A}_h N_{C,h,ij}$. The estimate of undercount in the population size of district j in state i is $N_{adj,ij} - N_{unadj,ij}$ and the estimate of the corresponding undercount rate is $(N_{adj,ij} - N_{unadj,ij})/N_{adj,ij}$. The estimate of undercount in the state population share for district j in state i is

$$\hat{U}_{ij} = \frac{N_{adj,ij}}{\sum_{k \in state\,i} N_{adj,ik}} - \frac{N_{unadj,ij}}{\sum_{k \in state\,i} N_{unadj,ik}}.$$

3. Sources of Error in Adjustments

The adjusted estimates are subject to a variety of possible sources of error: sampling error, data collection and survey operations error, missing data, error from exclusion of late census data and data with insufficient information for matching, contamination error, correlation bias, and synthetic estimation bias.

3.a. Sampling Error

Sampling error gives rise to random error, quantified by sampling variance, and to a systematic error known as ratio-estimator bias, which arises because even if X and Y are unbiased estimators, X/Y typically is biased. Random sampling error is reflected in the estimated covariance matrix of the $\hat{A}_h$'s. The covariance matrix is estimated by the Census Bureau's sampling-error software applied to the A.C.E. data. The software also provides estimates of ratio-estimator bias.

3.b. Data Collection and Survey Operations Error

Errors in reported data as well as errors in survey operations and processing of reported data cause errors in the components of the DSE. For example, processing errors can cause false matches and false nonmatches to be made between census enumerations and interview records. The estimates of biases due to error in data reporting and processing are based on evaluations of the 1990 post-enumeration survey (PES) because the evaluations of the A. C. E. will not be completed until well after April. For the 1990 PES, the Matching Error Study was used to estimate bias from P-sample matching error and E-sample processing error and the Evaluation Followup was used to estimate bias from P-sample fabrication and from P-sample and E-sample data collection error. For a detailed discussion of the definitions and estimation of the component errors see Appendices A and B.

In 2000, the search for matches occurred within all block-clusters and also in surrounding blocks for a sample of the cases with geocoding errors recorded in the E sample– a design called "Targeted Extended Search" (TES). The variance estimates for the A. C. E. account for the TES design. If the execution of the TES was flawed, biases will result. If, as seems to have occurred, the operational errors in the TES were comparable to those in the surrounding block search in 1990, the bias estimates in the total error model will roughly account for the biases from the TES; see Appendix A for further discussion.

The computation of $\hat{N}_{CE,h}$ requires census enumerations to be assigned to poststrata, and the computation of $\hat{N}_{P,h}/\hat{N}_{CP,h}$ requires P-sample enumerations to be assigned to poststrata. When the assignments are not made consistently for the two samples, error arises in the ratio $\hat{N}_{P,h}/\hat{N}_{CP,h}$. See Appendix G for further discussion.

3.c. Missing Data

A. C. E. data may be missing for a variety of reasons – some A.C.E. interviews fail to take place, some households provide incomplete data on questionnaire items, and in some cases the information for classification as a match or nonmatch is ambiguous. Methods are used to compensate for missing data, but they effectively assume that the match status for the case with missing data is equal on average to the status for cases that are similar except that they have complete data. Missing data on characteristics are imputed from otherwise similar cases with complete data. Nonresponse weighting adjustments are used to account for sampled but non-interviewed households. "Unresolved matches" are said to occur if the available data is inadequate to provide a determinate assignment of match or nonmatch, and in such cases a match status is imputed.

Information concerning bias from missing data comes from evaluations of the 1990 PES because the evaluations of the A. C. E. missing data will not be completed in time. Before describing how the assessment was done for 1990, we summarize by noting that estimated effect of bias was very minor, and consequently the analysis for 2000 assumes that the bias from missing data is negligible. The effect of imputation on variance of estimates is reflected in the estimates of sampling variance. See Appendix C for further details. The rationale for the treatment of bias is now described.

Although one can consider the range of effects on the DSE by considering extreme alternatives – e.g., all unresolved matches truly are matches or truly are non-matches – the range is too wide to be informative about the likely bias. The bias from the method used to compensate for missing data can in

principle be estimated from intensive followup of cases with missing data, but in practice the fraction completed by followup is too low. The Census Bureau analyzed the missing-data bias by looking at the changes in the DSE when alternative methods were used to compensate for missing data. The Bureau modeled the bias as a random effect whose variance was estimated from the changes observed in the DSE when alternative imputation methods were applied. Denote the latter variance for a poststratum in 1990 by $v_{1990\text{-}miss}$. On average, $v_{1990\text{-}miss}$ was 2 percent of the sampling variance. The corresponding variance component for 2000, say v_{miss}, would be estimated as $v_{1990\text{-}miss}$ times the square of the ratio of the fraction of unresolved match cases in 2000 to the fraction in 1990 times the square of the ratio of the population sizes in the poststratum for 2000 to that in 1990. The ratio of v_{miss} to the sampling variance of the population estimates for the evaluation poststrata, say r_{miss}, would then be calculated. The sampling variance-covariance matrix for the adjustment factors would then be multiplied by $1 + r_{miss}$ in order to reflect variance both from sampling and from choice of imputation method. The correlation matrix would be assumed to be the same as that for sampling error alone. For practical purposes, however, the 2% figure was small enough and the rate of unresolved matches was low enough that the effect on the accuracy was small enough to ignore.

3.d. Excluded-data Error

The DSE treats late census data as non-enumerations. Thus, duplicate enumerations among the late data do not contribute to census data but valid enumerations among the late data are treated as census misses and are estimated by the DSE. If the late census data were excluded from the entire adjustment process and estimation, no new source of error would be present. The adjusted estimates do partially incorporate late census data, by including them in $N_{C,h,ij}$ and $N_{C,h}$ but excluding them from the computation of $\hat{N}_h$. This use of late data affects the estimates for areas with disproportionately many or few late adds, with an effect that is similar to synthetic estimation error. In addition, the exclusion of late

census data from the E sample could bias the estimates at the poststratum level.

For insight into the cross-area effect, consider poststratum h in district j in state i and let $L_{h,ij}$ denote the number of late adds. If no late adds were used, which would be the consistent way to perform the estimation, the adjusted estimate would be $\hat{N}_h(N_{C,h,ij} - L_{h,ij})/(N_{C,h} - L_h)$. The estimation method, however, uses the estimate $\hat{N}_h N_{C,h,ij}/N_{C,h}$. The ratio of the former to the latter is $(1 - L_{h,ij}/N_{C,h,ij})/(1 - L_h/N_{C,h})$, which shows that the adjusted estimate is reduced for areas with disproportionately large numbers of late adds relative to what it would be if late adds were completely excluded from the computation.

There are two conditions that have to be met for the exclusion of the late adds from the processing of the A.C.E. not to bias the DSEs at the postratum level:

- The P sample covers the correct enumerations among the late adds at the same rate as other correct enumerations.
- The late adds occur in the E sample at the same rate as they occur in the census (excluding the imputations)

To see this, we will consider the conditions under which the inclusion of late adds in the DSE would not change its expected value. Define the following quantities for poststratum h, suppressing the subscript h.

L = late adds in poststratum h

$N_{C,L}$ = increase in correct enumerations in poststratum h if late adds are included in A.C.E. operations

$\hat{N}_{CE}$ = estimated population size for poststratum h from the E sample

$N_{CE,L}$ = expected increase in E-sample total in poststratum h if late adds are included in A.C.E. operations

$N_{CP,L}$ = expected increase in the matches in poststratum h if late adds are included in A.C.E. operations

One question is whether the following equality holds.

$$\hat{N}_{CE}\hat{N}_P/\hat{N}_{CP} = (\hat{N}_{CE} + N_{CE,L})\hat{N}_P/(\hat{N}_{CP} + N_{CP,L}).$$

It is readily seen that equality holds if and only if $N_{CE,L}/\hat{N}_{CE} = N_{CP,L}/\hat{N}_{CP}$, or $\hat{N}_{CP}/\hat{N}_{CE} = N_{CP,L}/N_{CE,L}$ Therefore, the percentage of correct enumerations that are matches among the late adds has to equal the percentage for the other correct enumerations. This means that the P sample has to cover the correct enumerations among the late adds at the same rate as other correct enumerations.

A second question concerns the weight adjustment in the E sample estimation. The weight adjustment will remain unchanged if

$$(N_C - I_C)/(\hat{N}_E + N_{EL}) = (N_C - I_C - L)/\hat{N}_E.$$

The equality holds if

$$1/(1 + N_{EL}/\hat{N}_E) = 1 - L/(N_C - I_C)$$

This equality holds approximately if $L/(N_C - I_C)$ and $N_{EL}/\hat{N}_E$ are small and equal. These ratios are equal when the late adds occur in the E sample at the same rate as they occur in the census minus the imputations.

It should be noted that a parallel analysis applies to the treatment of whole person imputations or cases with insufficient information for matching. We are *not* using data from 1990 to estimate bias from excluded late census data.

3.e. Contamination Error

Contamination occurs when the A.C.E. selection of a given block cluster alters the way the

census is conducted there and affects enumeration results, e.g, by increasing or decreasing erroneous enumerations or by increasing or decreasing coverage rates. Direct estimates of contamination error for Census 2000 will not be available until well after April. Analysis of the 1990 census and PES did not indicate the presence of contamination error. Therefore, the analysis will assume zero contamination error for Census 2000.

3.f. Correlation Bias

If there is variability of the enumeration probabilities for persons in the same poststratum or if there is a dependence between enumeration in the census and in the A.C.E. – e.g., people less likely to be enumerated in the census may also be less likely to be found in the A.C.E. – then correlation bias may arise. Correlation bias is most likely a source of downward bias in the DSE. Evidence of correlation bias in national estimates is provided by sex ratios (males to females) for adjusted numbers that are low relative to ratios derived from demographic analysis of data on births, deaths, and migration. The information from demographic analysis is insufficient to estimate correlation bias at the poststratum level, however, and alternative parametric models have been used to allocate correlation bias estimates for national age-race-sex groups down to poststrata. Estimates of correlation bias at the national level provided by demographic analysis information also account for possible error – if indeed it exists – from groups whose probabilities of enumeration are so low that the DSE will fail to account for them. The estimates of correlation bias based on sex ratios are affected by error in the demographic-analysis sex ratios and by possible other biases in the sex ratios in the DSE. The model selected for correlation bias is the "two-group" model, which assumes that the relative correlation bias is constant over male poststrata within age groups 18 to 29, 30 to 49, and 50 and over for Blacks and age groups 30 to 49 and 50 and over for Nonblacks. See Appendix D for further discussion.

3.g. Synthetic Estimation Bias

The adjustment methodology relies on a method called synthetic estimation to provide the same adjustment factor $\hat{A}_h$ for all enumerations in a given poststratum, regardless of whether the enumerations are from the same geographic area. Synthetic estimation bias arises when enumerations from different areas but in the same poststratum should have different adjustment factors. To assess synthetic estimation bias for a given area one needs to develop an estimate based on data from the area alone, which is rarely possible. Attempts to estimate synthetic estimation bias in undercount estimates from analysis of "artificial populations" or "surrogate" variables, whose geographic distributions are known, are unconvincing. Therefore, the estimates of Δ will be constructed without an allowance for synthetic estimation bias.

Omitting an allowance in $\hat{\Delta}$ for correlation bias can lead either to a positive or a negative bias in $\hat{\Delta}_i$ (and in $\hat{\Delta}$), depending on the sign of cross-area correlation between $E(\hat{U}_{ij})$ and the unmeasured bias from synthetic estimation in state i (Spencer 2001). For example, CAPE (1992) discussed some analyses indicating that the effect of ignoring synthetic estimation bias for the adjustments of the 1990 census was quite possibly to overstate the accuracy of both the census and the DSE and to favor the census relative to the DSE (i.e., to bias $\hat{\Delta}$ downward). Analyses will be conducted to assess the impact of synthetic estimation bias on the estimate of relative accuracy, $\hat{\Delta}$.

4. Total Error Modeling

The adjustment factor for poststratum h is subject to bias and variance. The variance-covariance matrix for sampling error in the adjustment factors $\hat{A}_h$ is estimated directly as, say, $\mathbf{V}_{samp}$, with h^{th} diagonal element $v_{samp,h}$. It reflects variance contributions from adjustment for missing data, as described in section 3.c and Appendix C.

The biases may be decomposed into the sum of ratio-estimator bias, data bias, contamination bias, excluded-data bias, and correlation bias.. We are treating contamination bias and excluded-data bias as zero. Thus, we will represent the expected value of the estimated adjustment vector as equal to the adjustment factor if there were not bias or random error, A_h, plus three bias terms:

$$\hat{A}_h = A_h + \text{ratio-estimator bias} + \text{data bias} + \text{correlation bias.}$$

As noted in section 3.a., ratio-estimator bias is estimated directly, say by $\hat{B}_{ratio,h}$ for poststratum h. We next consider the estimation of data bias.

To estimate data bias in the adjustment factor, let $\hat{B}_{CE-data,h}$ be an estimate of data bias in $\hat{N}_{CE,h}$ and let $\hat{B}_{P-data,h}$ and $\hat{B}_{CP-data,h}$ be estimates of data bias in $\hat{N}_{P,h}$ and $\hat{N}_{CP,h}$, respectively, and estimate the data bias in $\hat{A}_h$ by

$$\hat{B}_{data,h} = \hat{A}_h - \frac{\hat{N}_{CE,h} - \hat{B}_{CE-data,h}}{N_{C,h}} \times \frac{[\hat{N}_{P,h} - \hat{B}_{P-data,h}]}{\hat{N}_{CP,h} - \hat{B}_{CP-data,h}}.$$

The source of data for estimates of the data-bias components is the Matching Error Study and the Evaluation Followup for the 1990 PES. Although similar studies are being conducted for the A.C.E., the results will not be available by April. The data-bias estimates are developed by comparing two sets of data, the 1990 PES and the evaluation data (from the Matching Error Study and the Evaluation Followup), for the evaluation subsamples. Details are provided in Appendix B.

Correlation bias is estimated by $\hat{B}_{correl,h}$ for poststratum h; see Appendix D for details.

The estimate $\hat{B}_{ij}$ of the bias in the estimate of the population share of district j in state i, is estimated as

$$\frac{\tilde{N}_{ij}}{\sum_{k \in \text{state } i} \tilde{N}_{ik}} - \frac{N_{adj,ij}}{\sum_{k \in \text{state } i} N_{adj,ik}}$$

with $\tilde{N}_{ij} = \sum_h \tilde{A}_h N_{C,h,ij}$ and $\tilde{A}_h = \hat{A}_h - \hat{B}_{ratio,h} - \hat{B}_{data,h} - \hat{B}_{Correl,h}$. The variances $\hat{V}_{ij}$ are obtained by simulation, as described in Appendix F, where estimation of Δ is also described.

References

Bell, William (2001a) "Allowing for Correlation Bias in Total Error Model/Loss Function Analyses (using the two-group model) Draft dated January 30, 2001. Bureau of the Census.

Bell, William (2001b) "Estimating Correlation Bias in A.C.E. 2000 Estimates" Revised February 13, 2001. Bureau of the Census.

CAPE (1992) "Additional Research on Accuracy of Adjusted Versus Unadjusted 1990 Census Base for Use in Intercensal Estimates". Report of the Committee on Adjustment of Postcensal Estimates, Bureau of the Census, Department of Commerce, November 25, 1992.

Navarro, A. and Olson, D. (2001) "Accuracy and Coverage Evaluation Survey: Effect of Targeted Extended Search." DSSD Census 2000 Procedure and Operations Memorandum Series B-18. Bureau of the Census. Draft, dated February 16, 2001.

Parmer, Randall (1991) "Post Enumeration Survey Evaluation Project P11: Balancing Error Evaluation." 1990 Coverage Studies and Evaluation Memorandum Series #M-2. Bureau of the Census. Dated June 24, 1991.

Robinson, J. G. (2001) "Accuracy and Coverage Evaluation Survey: Demographic Analysis Results." DSSD Census 2000 Procedures and Operations Memorandum Series B-4. Bureau of the Census. Draft, dated February 16, 2001.

Singh, R. (1997) "Search Area Definition for the Dual System Estimation." Memorandum for Elizabeth A. Vacca, Decennial Statistical Studies Division, Bureau of the Census. Draft, dated October 31, 1997.

Spencer, B. D. (2001) A Method for Deciding Whether Adjustment of Census 2000 Improves Redistricting. Unpublished manuscript. Northwestern University. January 21, 2001

Spencer, B. D. (2000) Total Error Model for Census 2000: How Components of Error Can Be Estimated

from the Bureau's Planned Evaluation Studies, Final Report. May 11, 2000; rev. August 25, 2000. Cambridge: Abt Associates, Inc.

Appendix A: Use of 1990 Data to Estimate Nonsampling Errors in the A. C. E.

In some cases, data evaluating the nonsampling errors in the adjusted estimates based on the A.C.E. will not be available until well after April 2001. There are strong similarities between the adjusted estimates based on the 1990 PES and those based on the A.C.E., however, and the evaluations of the former will be used to provide information about the latter. The two designs differed in some respects. The following differences were taken into account in the re-analysis of the evaluation data. (i) The 2000 A.C.E. does not include any people in group quarters while the 1990 PES population included people in noninstitutional group quarters. (ii) In 2000, a non-mover reclassified to in-mover during the P-sample processing is removed from the P sample, whereas in 1990 such a reclassification from non-mover to mover remained in the P-sample. (iii) A person listed for the P sample but coded as removed in Before-Followup Matching could be reinstated in the P-sample After Followup Matching in 1990 but not in 2000. (iv) Different poststrata were used in 1990 and in 2000. P-sample and E-sample cases in the 1990 PES and evaluation subsamples were reassigned to poststrata based on the definitions in 2000 but using the data for 1990.

Certain differences did not require special treatment.

TES, In 2000, a search of the surrounding blocks was performed on a sample basis in the Targeted Extended Search rather than for all block clusters as in 1990. No specific treatment of this difference was required because the two designs yield the same expected values. Although the designs yield different variances, the variance estimates for the A.C.E. account for the TES design.

The estimation of the component errors for 2000 using data from the 1990 PES Evaluation reflects errors that may have occurred in the TES. The 1990 PES Evaluations included measurement of error in the surrounding block search. The Matching Error Study evaluated errors in the surrounding block search that occurred during the processing operations, and the Evaluation Followup evaluated errors that occurred during the data collection.

In 2000, approximately 3 million more matches than correct enumerations were found in surrounding blocks (Navarro and Olson, 2001) while in 1990, approximately 4.3 million more matches were found in surrounding blocks (Parmer, 1991). Analysis of the 2000 A.C.E. data show that the without the TES, the DSE at the national increases approximately 1.25% and the average increase for the seven age-sex groups is approximately 1.10%. Research with the 1990 data (Singh, 1997) showed that without the surrounding block search on any blocks the DSEs for the 357 poststrata increased 1.81% on average with the median of the increase as 1.50%. At the national level, the DSE was 1.94% higher.

In 2000, 3 million more matches than correct enumerations were found in surrounding blocks, while in 1990, 4.5 million more matches were found in surrounding blocks. Although data are not yet available, the conjecture is that the surrounding block search permits accounting for minor geocoding error in the P-sample listing and thereby improve the DSEs. By performing the surrounding block search, the P-sample members in housing units not in the sample block, but in adjacent blocks, find census matches. In 2000, the surrounding block search was conducted in blocks where there had been E sample geocoding errors. A reasonable assumption is that the P sample listing will have problems at many of the same places that the census had geocoding problems, and these are the blocks eligible for TES selection. The A.C.E. operations corrected some geocoding error by relisting block clusters that had a nonmatch rate of 45% or more, and by having a geocoding check on A.C.E. interviews were the interviewer changed the address. However, blocks where the nonmatch rate was lower than 45% did not have a field check of the geocoding, leaving the possibility of minor geocoding errors remaining in the P sample.

A reasonable conclusion is that the component errors cover error in the TES. The phenomena of the surrounding block search lowering the DSE and having more matches than correct enumerations have occurred in both 2000 and 1990. Also, a reasonable conclusion is that the surrounding block search, whether on a sample basis or everywhere, does not introduce bias, but rather reduces the bias that would

be present otherwise.

Treatment of Movers. Mover matching was done for out-movers in the 2000 A.C.E. instead of the in-movers as in the 1990 PES. It is not known that either method will lead to more accurate DSEs than the other. Accordingly, we assume that the error arising from treatment of movers in the A.C.E. was similar to the error from treatment of movers in the PES.

We use several sources of data to estimate the first two moments of the component errors for the 2000 A.C.E. Some errors we estimate using the 2000 demographic analysis and the 2000 A.C.E. itself. Since the evaluations of data collection and processing errors in the 2000 A.C.E. are not available yet, we base estimates of these component errors on the data collected in the evaluations of the 1990 PES. For our analysis, we use 16 evaluation poststrata, which are aggregates of A.C.E. poststrata. These evaluation poststrata, shown in Table 1, are formed by grouping poststrata based on race/ethnicity, tenure, population density, region, type of enumeration, and census mail response rate. The minorities are Blacks, Hispanics, Asians, and American Indians. Table 2 shows the source of the data used to estimate the components of error included in the 2000 A.C.E. and the 1990 PES. Before giving a detailed discussion in Appendix B, we provide an overview.

We used data from the 1990 PES Evaluation Master Variance File to estimate the data collection and processing errors for the 2000 A.C.E. Calculating estimates with this file meant using characteristics from the 1990 Census to assign each record to a 2000 A.C.E. poststrata. This way, we have comparability by applying component errors for blocks with a high mail response rate in the 1990 Census to blocks with a high mail response rate to the 2000 Census. Table 1 shows the number of records in the 1990 PES Evaluations Master Variance File for each of the 2000 A.C.E. evaluation poststrata. Table 19 shows the estimated undercount rates for the 1990 Census and the 2000 Census for the 2000 A.C.E.

evaluation poststrata.

Since the 1990 Evaluation Sample is not large enough to support reliable direct estimates for the 2000 A.C.E. poststrata, we first compute direct estimates for the 16 evaluation poststrata and then form model-based estimates for the poststrata. We use synthetic estimation methodology for the model-based estimation. For sampling variance, imputation variance, correlation bias, and ratio estimator bias, we use direct estimates for the poststrata.

The synthetic estimation has two phases. First, we apply the synthetic estimation to the estimates of the gross component errors to distribute the them to the seven age-sex groups within each evaluation poststratum, called the intermediate poststrata. After generating the bias estimates from the total error simulation for the 112 intermediate poststrata, we distribute the bias estimates among the poststrata within an intermediate poststratum two alternative ways, proportional to the DSE, denoted by GROSDSE, and proportional to the absolute value of the estimate of net undercount, denoted by GROSUC.

The synthetic estimation method has some advantages. The ratios of the component errors between any two age-sex groups within each minority (or nonminority) evaluation poststratum then equal the ratios for minorities (or nonminorities) at the national level. For distribution proportional to the DSE (GROSDSE), the relative bias in the DSEs for the A.C.E. poststrata equals the relative bias in the DSE of their intermediate poststratum. For distribution proportional to the absolute net undercount (GROSUC), the poststratum with the largest absolute net undercount has the largest portion of the bias. A possible drawback to the synthetic method is that it may not fully reflect differences in biases across poststrata.

To describe the estimation for a component error, let u^+ and u^- denote the means of the positive and negative gross errors for the component. For example, a positive gross error is the number of misclassifications into a category of interest (e.g., false matches) and the negative gross error is the number of misclassifications out of the category (e.g., false nonmatches). Let u denote the mean of the

net error, $u = u^{(+)} - u^{(-)}$. To calculate the synthetic estimates we first derive the direct estimate $u_j^{(+)}$ of the mean of a positive gross error component in evaluation poststratum j, and we derive an estimate of the sampling variance of the estimate, $\sigma_j^{(+)2}$. We estimate the negative gross error component and its variance in the same way. Tables 3 through 18 show the moments of the gross and net component errors for the 16 evaluation poststrata. We directly estimate the sampling covariance $\sigma_{juv}^{(+-)}$between the positive gross error component of u and the positive gross error component of v in evaluation poststratum j, the sampling covariance $\sigma_{juv}^{(+-)}$between the positive gross error component of u and the negative gross error component of v in evaluation poststratum j, etc. We then split the evaluation poststrata into two major groups, minority and nonminority. For each of the seven age-sex groups in each of the two major groups of evaluation poststrata, we derive direct estimates of each gross error component $u_{[i,j]}^{-}$and $u_{[i,j]}^{-}$, with [i,j] referring to an age-sex group in the minority or nonminority group of evaluation poststrata to which evaluation poststratum j belongs. Let major(j) denote the major evaluation poststratum group (minority or nonminority) to which evaluation poststratum j belongs and let $u_{[major(j)]}^{+} = \sum u_{[i,j]}^{-}$ with summation occurring over age-sex groups within major(j). We then estimate an error component and its variance for an intermediate poststratum by $u_j^{+}(u_{[i,j]}^{+}/u_{[major(j)]}^{+})$ and $\sigma_j^{(+)2}(u_{[i,j]}^{-}/u_{[major(j)]}^{+})^2$ where j denotes the corresponding evaluation poststratum and i denotes the corresponding minority or non minority age-sex group. The covariance between the kth and mth error components in an intermediate poststratum is estimated by $\sigma_{juv}^{+-}(u_{[i,j]}^{+}/u_{[major(j)]}^{-})(v_{[i,j]}^{-}/v_{[major(j)]}^{-})$.

A few comments on using 1990 evaluation data to estimate data biases for the A.C.E. are in order. Use of the 1990 bias estimates (adjusted for differences in poststrata and changes in population sizes) leads to roughly the same dispersion of bias estimates across poststrata but does not necessarily put the bias estimates in the correct poststrata unless the actual poststratum-level data biases are stable across the two censuses. Figure 5b shows no association between estimated biases and estimated undercount rates for 1990 at the evaluation poststratum level, however, and a similar lack of association for 2000.

This similarity of pattern is desirable. More penetrating tests of the validity of the use of 1990 bias estimates for the A.C.E. can occur when the evaluations of the A.C.E. are completed and direct estimates of data bias can be obtained.

Appendix B: Estimation of Data Bias

This appendix contains the definitions of component errors and formulas for direct estimates of them using the 1990 PES Evaluations Master Variance File. Appendix A describes how the synthetic estimation uses these direct estimates in the total error model. Here we suppress the subscript h which would indicate poststratum h. The error component definitions are

$\hat{B}_{CP-data} = m_m + m_a + m_f$

$\hat{B}_{P-data} = n_{Pm} + n_{Pa} + n_{Pf}$

$\hat{B}_{CE-data} = c_o + c_{resp}$

where the net sources of nonsampling error are defined by:

n_{Pm} = mean of matching error component of $\hat{N}_P$

m_m = mean of matching error component of $\hat{N}_{CP}$

n_{Pa} = mean of data collection error component of $\hat{N}_P$

m_a = mean of data collection error component of $\hat{N}_{CP}$

n_{Pf} = mean of fabrication error component of $\hat{N}_P$

m_f = mean of fabrication error component of $\hat{N}_{CP}$

c_o = mean of office processing error component of $\hat{N}_{CE}$

c_{resp} = mean of data collection (respondent) error component of $\hat{N}_{CE}$.

Each error component source can be defined by the difference of the expected gross errors:

$n_{Pm} = n_{Pmp} - n_{Pmn}$

$m_m = m_{mp} - m_{mn}$

$n_{Pa} = n_{Pap} - n_{Pan}$

$m_a = m_{ap} - m_{an}$

$n_{Pf} = n_{Pfp} - n_{Pfn}$

$m_f = m_{fp} - m_{fn}$

$c_o = c_{op} - c_{on}$

$c_{resp} = c_{respp} - c_{respn}$

Definition of E-sample office processing errors, c_{op} and c_{on} in 1990 and for 2000

For 1990, the positive component of the gross error for office processing error (c_{op}) is defined by the weighted number of cases changed from correct (CE) to erroneous (EE) plus the number of cases changed from CE to U (unresolved) multiplied by the rate of change among the cases that were originally CE that were resolved in the rematch. The second term is a simple imputation for the unresolved cases. The negative component (c_{on}) is defined analogously. Both have an adjustment of the ratio of the nonimputed and data defined persons (DDEFPER) to the weighted E-sample total (WTEPER). The DSE includes the same ratio adjustment to account for the sampling error in the E-sample total. Specifically, we have

$$c_{op} = [CEtoEEco + CEtoUco \times (CEtoEEco/(CE - CEtoUco))] \times (DDEFPER/WTEPER)$$

$$c_{on} = [EEtoCEco + EEtoUco \times (EEtoCEco/((EE - EEtoUco))] \times (DDEFPER/WTEPER)$$

where

CEtoEEco = number of cases changed from correct (CE) to erroneous (EE)

CEtoUco = number of cases changed from CE to U (unresolved)

CE = number of cases coded correct in production

EEtoCEco = number of cases changed from erroneous (EE) to correct (CE)

EEtoUco = number of cases changed from EE to U (unresolved)

EE = number of cases coded erroneous in production.

For 2000, we first must adjust the 1990 estimates of c_{op} and c_{on} to account for the change in the definition of the population to be only those people living in housing units. The 1990 PES also included the people living in noninstitutional group quarters (NI). All the definitions above will have the suffix 'hu' which means that only those people enumerated in housing units will be included. Some of the people in housing units match to people in group quarters, but we will not account for that.

cophu = [CetoEEcohu + CEtoUcohu×(CEtoEEcohu/(CEhu- CEtoUcohu))] ×(DDEFPERhu/WTEPERhu)

conhu = [EEtoCEcohu+EEtoUcohu×(EEtoCEcohu/(EEhu- EEtoCEcohu))] ×(DDEFPERhu/WTEPERhu)

Next, we will multiply the 1990 estimate of the number of gross errors by the ratio adjustment defined by the number of correct enumerations in 2000 (CE2) divided by the number of correct enumerations in 1990 (CEhu). There is also a ratio adjustment to account for the difference in the ratio of the nonimputed and data defined persons (DDEFPER2) to the E-sample total (WTEPER2) between 1990 and 2000.

For variance estimation this ratio adjustment will be treated as a constant with the assumption that the contribution of the adjustment to the variance is the same in 2000 as it was in 1990. The ratio adjustment CE2/CE also is treated as a constant. These assumptions permit using 1990 VPLX programs in the variance computations. For 2000 we set $c_{op} = c_{op2} - c_{on2}$, with

c_{op2} = cophu×CE2/CEhu×(DDEFPER2/WTEPER2)/(DDEFPERhu/WTEPERhu)

c_{on2} = conhu×EE2/EEhu×(DDEFPER2/WTEPER2)/(DDEFPERhu/WTEPERhu)

Definition of E-sample data collection errors c_{respp} and c_{respn} in 1990 and 2000

For 1990, the positive component of the gross error for data collection error (crespp) is defined by the weighted number of cases changed from correct (CE) to erroneous (EE) plus the number of cases changed from CE to U (unresolved) multiplied by the rate of change among the cases that were originally CE that were resolved in the rematch. The second term is a simple imputation for the unresolved cases. Both have an adjustment of the ratio of the nonimputed and data defined persons (DDEFPER) to the E-sample total (WTEPER). As stated above, the DES includes the same ratio adjustment to account for the sampling error in the E-sample total. We have

$$c_{respp} = [CEtoEEcr + CEtoUcr \times (CEtoEEcr/(CE - CEtoUcr))] \times (DDEFPER/WTEPER)$$

$$c_{respn} = [EEtoCEcr + EEtoUcr \times (EEtoCEcr/(EE - EEtoUcr))] \times (DDEFPER/WTEPER)$$

where

CEtoEEcr = number of cases changed from correct (CE) to erroneous (EE)

CEtoUcr = number of cases changed from CE to U (unresolved)

CE = number of cases coded correct in production

EEtoCEcr = number of cases changed from erroneous (EE) to correct (CE)

EEtoUcr = number of cases changed from EE to U (unresolved)

EE = number of cases coded erroneous in production.

For 2000, we also need to adjust the population so that people in noninstitutional group quarters are not included.

crespphu=[CEtoEEcrhu+CEtoUcrhu×(CEtoEEcrhu/(CEhu- CetoUcrhu))] ×(DDEFPERhu/WTEPERhu)

crespnhu=[EEtoCEcrhu+EEtoUcrhu×(EEtoCEcrhu/(EEhu- EEtoUcrhu))]×(DDEFPERhu/WTEPERhu)

Next, as for the operations error, we will multiply the 1990 estimate of the number of gross errors by the ratio adjustment defined by the number of correct enumerations in 2000 (CE2) divided by the number of correct enumerations in 1990 (CEhu). This ratio adjustment accounts for the different sample size and CE rates. There is also a ratio adjustment to account for the difference in the ratio of the nonimputed and data defined persons (DDEFPER2) to the E-sample total (WTEPER2) between 1990 and 2000.

For variance estimation this ratio adjustment will be treated as a constant with the assumption that the contribution of the adjustment to the variance is the same in 2000 as it was in 1990. The ratio adjustment CE2/CE also is treated as a constant. These assumptions permit using 1990 VPLX programs in the variance computations. For 2000 we set $c_{resp} = c_{respp2} - c_{respn2}$, with

c_{respp2}= crespphu×CE2/CEhu×(DDEFPER2/WTEPER2)/(DDEFPERhu/WTEPERhu)

c_{respn2}= crespnhu×EE2/EEhu×(DDEFPER2/WTEPER2)/(DDEFPERhu/WTEPERhu)

Definition of P-sample fabrication errors n_{Pf} and m_f in 1990 and for 2000

P-sample fabrication errors n_{Pf} and m_f each have only a negative component. The error n_{Pf} accounts for the tendency of the average size of fabricated households found in the Evaluation Followup (AVEHHPF) to be smaller than overall average household size (AVHHPES). The error m_f accounts for the inability to match fabricated persons. All factors are weighted in the following formula:

n_{Pf} = - (estimated number fabricated)×((AVHHSIZE/AVHHPF) – 1)

m_f = - (estimated number fabricated)×MATCHRATE + nPf×MATCHRATE.

For 2000, we must adjust the 1990 P-sample population to include only those people living in housing units so that it will be comparable to the population for the 2000 A.C.E. Also, we adjust the match rate so that it is based only on the nonmovers in housing units.

nPfhu = - (estimated number fabricated)×((AVHHSIZE/AVHHPF) – 1)

mfhu= - (estimated number fabricated)×MATCHRATEnmhu + nPfhu×MATCHRATEnmhu

Next, we will multiply the 1990 estimate of the number of error by the ratio adjustment defined by the size of the P-sample in 2000 (WTPPER2) divided by the size of the P-sample in 1990 (WTPPERhu). This ratio adjustment accounts for the different sample size. For variance estimation, this ratio will be treated as a constant.

For 2000 we set $m_f = m_{f2}$ and $n_{Pf} = n_{Pf2}$, with

n_{Pf2} = nPfhu × (WTPPER2/WTPPERhu)

m_{f2} = mfhu × (WTPPER2/WTPPERhu).

Definition of P-sample gross data collection errors in 1990 and 2000

n_{Pap} and n_{Pan}

For 1990, the positive component of the gross error for P-sample population size due to data collection error (nPap) is defined by the number of cases changed to another evaluation poststratum (MOVOUTa) plus the number of cases changed to out of scope (INtoOUTa). The negative component (nPan) is defined analogously. All factors are weighted in the following formulas:

n_{Pap} = MOVOUTa + INtoOUTa

n_{Pan} = MOVINa + OUTtoINa.

For 2000, we must adjust the 1990 P-sample population to include only those people living in housing units so that it will be comparable to the population for the 2000 A.C.E. We retain this basic definition of the error component although the processing for the 2000 A.C.E. is different because we think it is the best way to represent the net error. The 2000 treatment of movers in the P sample means that no cases are changing poststratum, and therefore, no one will be changed from an in-mover to a non-mover. Also, no one will be added to the P sample for other reasons. Also, we will multiply the 1990 estimate of the number of errors by the ratio adjustment defined by the size of the P-sample in 2000 (WTPPER2) divided by the size of the P-sample in 1990 (WTPPERhu). This ratio adjustment accounts for the different sample size. For variance estimation, this ratio will be treated as a constant.

For 2000, n_{Pa} is estimated by $n_{Pap2} - n_{Pan2}$, with

n_{Pap2} = nPaphu × (WTPPER2/WTPPERhu)

n_{Pan2} = nPanhu × (WTPPER2/WTPPERhu).

$\underline{m_{ap} \text{ and } m_{an}}$

For 1990, the positive component of the gross error for P-sample matches due to data collection error (m_{ap}) is defined by the sum of

- number of matches changed to nonmatches (MtoN)
- number of cases changed from M to U (unresolved) multiplied by the observed change rate among the cases that were originally CE that were resolved in the rematch as a simple imputation for the unresolved cases.
- number of cases changed to another evaluation poststratum that were matches (MOVOUTMa)
- number of cases changed to out of scope that were matches (INtoOUTMa).

The negative component (m_{an}) is defined analogously. All factors in the following formulas are weighted.

m_{ap} = MtoNMa + MtoUa×(MtoNMa/((M-MtoUa)) + MOVOUTMa + INtoOUTMa

m_{an} = NMtoMa + NMtoUa×(NMtoMa/(NM- NMtoUa)) + MOVINMa +OUTtoINMa

For 2000, we will use the same basic formulas because we think they are the best way to estimate the net errors as we did for the estimation of the P-sample population size. Although the 1990 PES included in-movers in the P-sample population while the 2000 A.C.E. instead includes the out-movers, we will assume that the error rate due to data collection is comparable.

maphu = MtoNMahu + MtoUahu×(MtoNMahu/(Mhu - Mto Uahu)) + MOVOUTMahu + INtoOUTMahu

manhu = NMtoMahu + NMtoUahu×(NMtoMahu/(NMhu - NMtoUahu)) + MOVINMahu + OUTtoINMahu

Next, we will multiply the 1990 estimate of the positive errors by the ratio adjustment defined by the number of matches in 2000 (M2) divided by the number of matches in 1990 (Mhu). This ratio adjustment accounts for the different sample size and match rates. The negative gross error estimate will be adjusted analogously.

The ratio adjustments M2/M and NM2/NM are treated as constants. For 2000, m_a is estimated by m_{ap2} - m_{an2}, with

m_{ap2} = maphu×M2/Mhu

m_{an2} = manhu×NM2/NMhu.

Definition of P-sample gross matching errors in 1990 and 2000

n_{Pmp} and n_{Pmn}

For 1990, the positive component of the gross error for P-sample population size due matching error (nPmp) is defined by the number of cases changed to another evaluation poststratum (MOVOUTm) plus the number of cases changed to out of scope (INtoOUTm). The negative component (nPan) is defined analogously. All factors in the following formulas are weighted.

n_{Pmp} = MOVOUTm + INtoOUTm

n_{Pmn} = MOVINm + OUTtoINm

For 2000, we will continue to use the same basic formulas as we did for n_{Pap} and n_{Pan} because it is the best way to represent the net error in light of the 1990 PES and 2000 A.C.E. having different treatments of movers. Also, we will multiply the 1990 estimate of the number of errors by the ratio adjustment defined by the size of the P-sample in 2000 (WTPPER2) divided by the size of the P-sample in 1990 (WTPPERhu). This ratio adjustment accounts for the different sample size. For variance estimation, this ratio will be treated as a constant. For 2000, n_{Pm} is estimated by $n_{Pmp2} - n_{Pmn2}$, with

n_{Pmp2} = nPmphu × (WTPPER2/WTPPERhu)

n_{Pmn2} = nPmnhu × (WTPPER2/WTPPERhu).

m_{ap} and m_{an}

For 1990, the positive component of the gross error for P-sample matches due to matching error (m_{mp}) is defined by the sum of

- number of matches changed to nonmatches (MtoNm)
- number of cases changed from M to U (unresolved) multiplied by the observed change rate among the cases that were originally CE that were resolved in the rematch as a simple imputation for the unresolved cases.
- number of cases changed to another evaluation poststratum that were matches (MOVOUTMm)
- number of cases changed to out of scope that were matches (INtoOUTMm).

The negative component (m_{mn}) is defined analogously. All factors in the following formulas are weighted.

m_{mp} = MtoNMm + MtoUm×(MtoNMm/(M-MtoUm)) + MOVOUTMm + INtoOUTMm

m_{mn} = NMtoMm + NMtoUm×(NMtoMm/(NM-NmtoUm)) + MOVINMm +OUTtoINMm

Next, we will continue to use these formulas because they give the best representation of the net error even though the treatment of movers is different in the 1990 PES and the 2000 A.C.E.

mmphu = MtoNMmhu + MtoUmhu×(MtoNMmhu/(Mhu- MtoUmhu)) + MOVOUTMmhu + INtoOUTMmhu

mmnhu = NMtoMmhu + NMtoUmhu×(NMtoMmhu/(NMhu-NMtoUmhu)+ MOVINMmhu +OUTtoINMmhu

Finally, we will multiply the 1990 estimate of the positive errors by the ratio adjustment defined by the

number of matches in 2000 (M2) divided by the number of matches in 1990 (Mhu). This ratio adjustment accounts for the different sample size and match rates. The negative gross error estimate will be adjusted analogously. The ratio adjustments M2/Mhu and NM2/NMhu are treated as constants. For 2000, m_{mp} is estimated by $m_{mp2} - m_{mn2}$, with

m_{mp2} = mmphu×M2/Mhu

m_{mn2} = mmnhu×NM2/NMhu.

Appendix C: Total Variance

We use the total variance of the undercount rate when forming confidence intervals. The total variance, V, of the estimated undercount rate is the sum of 3 terms:

$$V = S^2 + V_{ns} + V_M$$

with

S^2 = sampling variance

V_{ns} = variance of the nonsampling bias

V_M = variance due to imputation

$= V_{RA} + V_B + V_I$

V_{RA} = variance due to the imputation model selection

V_B = variance due to the model parameter estimation

V_I = within- person imputation variance.

The estimate of variance of the nonsampling bias is a byproduct of the total error simulations. The imputation variance components due to parameter estimation and within person estimation are included in the sampling error estimates, leaving the variance due to model selection. To estimate this component, we will use the results of the evaluation of imputation error in the 1990 PES, which estimated the imputation variance components separately. The variance due to model selection was 2% of the sampling error on the average in 1990.

Therefore, we will estimate V_{RA} by $0.02*S_{90}^2$ multiplied by the square of the product of the ratio of the 2000 DSE to the 1990 DSE and the ratio of the percentage of match codes imputed in 2000(P_{00}) to the percentage of match codes imputed in 1990 (P_{90}). This percentage is the sum of the E-sample imputations and the P-sample imputations. Also, we will use the S_{90}^2 from the 1992 version of the 1990 PES estimates instead of S_{00}^2 from the 2000 A.C.E. to obtain the appropriate order of magnitude.

Appendix D: Correlation Bias

The assumptions and model underlying the measurement of correlation bias are discussed in detail in a paper by Bell (2001a), but we will describe them briefly here. Although there are several models for how correlation bias is distributed, our main model is the "two-group" model. We rely on the basic assumptions listed below for the estimation of correlation bias and in addition, we conduct sensitivity analyses to assess the impact of these assumptions.

- The ratio of males to females measured in demographic analysis is more reliable for the two racial groups, Black and Nonblack, than the A.C.E., the exception of the sex ratios for the Nonblacks aged 18 to 29, which is discussed further below.
- There is no correlation bias present in the A.C.E. estimates for females.
- The relative correlation bias is equal across all A.C.E. poststrata within an age-race category.
- The relative impact of other nonsampling errors is equal for males and females at the national level.

The assumption with the two-group model of the relative correlation bias being equal across poststrata within an age-sex category has the advantage of permitting the estimation of correlation bias through a multiplicative factor applied to the corrected DSE. Even more important, an unbiased estimate of the factor is available under assumption that the relative impact of the other nonsampling errors is equal for males and females without actually having to estimate the nonsampling errors.

The sex ratios from demographic analysis in 1990 and 2000 along with those from the 1990 PES and 2000 A.C.E. are shown below. (Bell,2001b)

Ratios of Males to Females from DA and 2000 A.C.E

Age	Black		Nonblack	
	A.C.E.	DA	A.C.E.	DA
18-29	0.84	0.90	1.05	1.03
30-49	0.82	0.89	1.00	1.00
50+	0.72	0.75	0.85	0.86

Ratios of Males to Females from DA and 1990 PES

Age	Black		Nonblack	
	PES	DA	PES	DA
18-29	0.83	0.90	1.02	1.02
30-49	0.84	0.91	0.99	1.01
50+	0.72	0.78	0.81	0.82

The comparison between sex ratios based on demographic analysis and dual system estimation in both 1990 and 2000 is instructive. The sex ratios for Blacks from the two estimation methodologies are similar in 1990 and 2000. However, the same cannot be said for the Nonblacks. While the sex ratios for the 30-49 and 50+ age groups are larger in 2000 for both methods, the relative difference in the two sets is similar. However, the same cannot be said for the 18-29 age group. Surprisingly, the sex ratio from demographic analysis is lower than the sex ratio from the A.C.E. The relative percentage of correlation bias in the 2000 A.C.E. and the 1990 PES based on the sex ratios shown above follows (Bell, 2001b):

Relative Correlation Bias Estimates

2000 A.C.E (%)

Age	Black	Nonblack
18-29	-7.4	2.0
30-49	-8.0	-0.2
50+	-4.8	-0.9

Relative Correlation Bias Estimates

1990 PES (%)

Age	Black	Nonblack
18-29	-8.0	-0.3
30-49	-7.7	-1.6
50+	-8 2	-1.2

Since the estimates of the relative correlation bias are questionable, we perform the total error analyses under four assumptions for correlation bias with the estimates based on the sex ratios from demographic analysis:

- No correlation bias
- Correlation bias is present for Black males but not for Nonblack males
- Correlation bias is present for all males with the exception of Nonblacks in the 18-29 age group.
- Correlation bias is present for all males, including the 2% overcount of Nonblack males in the 18-29 age group.

When reviewing the estimates of correlation bias based on the sex ratios from demographic analysis, our judgment is that the most realistic scenario uses these estimates with the exception of the estimates for the Nonblack males 18 to 29 years of age. Including estimates of correlation bias in the total error analysis is appropriate because no other evidence or theory suggests that the problem of correlation bias has been solved for dual system estimation of census coverage error.

Using the estimates of correlation bias based on the demographic analysis sex ratios for all but the 18-29 Nonblack males is reasonable because the rates of correlation bias for Blacks in the 2000 A.C.E. are similar to those for Blacks in 1990. As for the estimates for Nonblack males, some evidence exists to support the conjecture that demographic analysis estimates probably underestimate the amount of illegal immigration among the Nonblack population, particularly for Hispanics. Therefore, we do not use the estimates of a 2 percent overcount for 18 to 29 Nonblack males because this is the age group where the heaviest illegal immigration occurs. We concede that the estimates of the correlation bias for the 30-49 Nonblack males and 50+ Nonblack males may underestimate the level of correlation bias, but prefer underestimating to assuming that no correlation bias exists in these groups. For a discussion of demographic analysis, see Robinson (2001). Alternative assumptions to use in studying the effect of correlation bias include assuming the correlation bias estimated for Blacks also holds for other minorities and varying the assumptions about the distribution of correlation bias among the poststrata.

Appendix E: Alternative Loss Functions

Loss functions are summary measures of error in estimates. The estimates considered here are adjusted or unadjusted measures of population size or shares of population. Several loss functions for measuring error in population estimates have been considered. The loss functions are all based on squared differences between the estimate and the quantity being estimated, or what we will call the squared errors. The various loss functions differ from one another in how weighting factors are applied to summarize the results across states and across substate areas, such as Congressional districts.

In random sampling, the actual squared errors typically are not known. It is customary therefore to consider the *expected* squared error, which can be estimated. Another name for expected squared error is mean squared error, or MSE. Thus, if a sample-based estimate is unbiased, its expected squared error (or MSE) is equal to its variance. The expected values of the loss functions considered here are shown symbolically below.

Notation

i	state or District of Columbia $\quad 1 \le i \le 51$
n_i	number of districts (or substate areas) within state i
ij	district j (or substate area j) within state i $\quad 1 \le j \le n_i$
$\tilde{P}_{ij}$	measure of population size of district j (or substate area j) within state i
$\tilde{P}_i$	measure of population size of state i
$MSE_{share,ij}$	mean squared error of the estimated proportion (or share) of state population i that is in district j (or subarea j)
$MSE_{level,ij}$	mean squared error of the estimated size of district j (or subarea j) in state i

Expected Values of Loss Functions for Estimates of Shares

1. $\sum_{i=1}^{51} \sum_{j=1}^{n_i} MSE_{share,ij}$

2. $\sum_{i=1}^{51}\sum_{j=1}^{n_i}\frac{MSE_{share,ij}}{\tilde{P}_{ij}/\tilde{P}_i}$

3. $\sum_{i=1}^{51}\sum_{j=1}^{n_i}\frac{MSE_{share,ij}}{(\tilde{P}_{ij}/\tilde{P}_i)^2}$

4. $\sum_{i=1}^{51}\tilde{P}_i^2\sum_{j=1}^{n_i}MSE_{share,ij}$

Expected Values of Loss Functions for Estimates of Population Size

5. $\sum_{i=1}^{51}\sum_{j=1}^{n_i}MSE_{level,ij}$

6. $\sum_{i=1}^{51}\sum_{j=1}^{n_i}\frac{MSE_{level,ij}}{\tilde{P}_{ij}}$

7. $\sum_{i=1}^{51}\sum_{j=1}^{n_i}\frac{MSE_{level,ij}}{\tilde{P}_{ij}^2}$

Note that the measures of size $\tilde{P}_{ij}$ and $\tilde{P}_i$ should not depend on which estimate is being evaluated, but should be chosen conventionally. It is recommended that the unadjusted census count be used for $\tilde{P}_{ij}$ and $\tilde{P}_i$.

Alternative substate areas than districts can be used, such as counties or even subsets of counties. If the substate areas do not comprise the entire state, then the definition of share could be based either on the state population or on the sum of population sizes of the substate areas. Whichever definition is used, the measure of size $\tilde{P}_i$ should be chosen commensurately, either as the state population or as the sum of $\tilde{P}_{ij}$ computed over the substate areas j within state i.

Recall that the measure of improvement for accuracy in redistricting is

$$\Delta = n^{-1} \sum_{1 \le i \le 51} \tilde{P}_i^2 \sum_{1 \le j \le n_i} (MSE_{unadj,ij} - MSE_{adj,ij})$$

with n_i the number of congressional districts in state i (where for these purposes we include the District of Columbia as a state), $n = \sum_{1 \le i \le 51} n_i$ the total number of districts (or representatives), $\tilde{P}_i$ the size of state i, and $MSE_{unadj,ij}$ and $MSE_{adj,ij}$ the mean squared errors of the estimated proportion (or share) of state population i that is in district j based on unadjusted and on adjusted estimates, respectively.

To help interpret this, note that it corresponds to a loss function equal to 1/n times loss function number 4,

$$n^{-1} \sum_{1 \le i \le 51} \sum_{1 \le j \le n_i} \tilde{P}_i^2 MSE_{est,ij},$$

with $MSE_{est,ij}$ the mean squared error in the estimate of state i's population share held by district j. The quantity $\tilde{P}_i^2 MSE_{est,ij}$ can be interpreted as the mean squared error in the estimated size of the district j in state i if the estimated total for state i were equal to the true total and they were both equal to $\tilde{P}_i$.

The interpretations of the loss function or measure of improvement Δ may be facilitated if they are multiplied by an appropriate positive constant. We suggest that the constant be taken to be the reciprocal of the square of the average district size, or $1/(n^{-1} \sum_{1 \le i \le 51} \tilde{P}_i)^2$. This leads to a scaled loss function,

$$\frac{n^{-1} \sum_{1 \le i \le 51} \sum_{1 \le j \le n_i} \tilde{P}_i^2 MSE_{est,ij}}{(n^{-1} \sum_{1 \le i \le 51} \tilde{P}_i)^2}$$

and a scaled measure of improvement,

$$\Delta_{scaled} = \frac{n^{-1} \sum_{1 \le i \le 51} \tilde{P}_i^2 \sum_{1 \le j \le n_i} (\dot{MSE}_{unadj,ij} - \dot{MSE}_{adj,ij})}{(n^{-1} \sum_{1 \le i \le 51} \tilde{P}_i)^2}.$$

Appendix F: Loss Function Calculation

This appendix summarizes the way calculations of loss functions were performed and explains the logic behind the calculations. The logic of the analysis is fairly straightforward, but is easily lost in the trees. To explain it we use some simple notation, which will be replaced by more complex notation when the details are described, below. (Also see section 1, above.) Let C denote the census estimate, A the adjusted estimate, and B an estimate of bias in the adjusted estimate. Let V_A denote an estimate of variance of A and let V_B denote an estimate of variance of B; we assume A and B have negligible correlation. To estimate the mean squared error (MSE) of C and A we construct a "target" estimate, T, defined as T = A - B. If T had zero variance, we could estimate the MSEs by $(C - T)^2$ and $(A - T)^2$. The variance of T is approximately $V_A + V_B$, however, and so we estimate the MSE of C by

$$(C - T)^2 - (V_A + V_B) \tag{1}$$

and we estimate the MSE of A by

$$B^2 + V_A - V_B. \tag{2}$$

The excess MSE of C relative to the MSE of A is estimated by $(C - T)^2 - B^2 - 2V_A$; observe that the specification for V_B does not affect point estimates of the difference in the MSEs.

The variances are calculated by means of replicates. The basis for the calculation of adjusted estimates and targets consists of (i) the vector of adjustment factors for poststrata, (ii) the estimated covariance matrix of the adjustment factors, (iii) the vector of estimated biases of the adjustment factors, and (iv) the estimated covariance matrix of the estimated biases. The vectors of replicates are constructed by random sampling from a multivariate normal distribution with covariance matrix equal to the estimated covariance matrix of (i) or (iii), as the case may be. To estimate the variance of a function of (i) or (iii), we calculate the function for each replicate and use the empirical variance among the

calculated values.

Notation

Subscript h ($1 \le h \le H$) will refer to poststrata and the subscript i ($1 \le i \le I$) will refer to general areas such as states, counties, cities, congressional districts, etc. The subscripts q ($1 \le q \le Q$) and s ($1 \le s \le S$) will be used to denote replicates. The replicates are constructed so that their empirical variance over q provides an estimate of variance due to random sampling in the DSE (see $\hat{V}_{ahℓ}$, below) and their empirical variance over s provides an estimate of variance due to random sampling in the evaluation studies for estimating bias in the DSE (see $\hat{V}_{thℓ}$, below) ; details are provided below. A "+" in place of a subscript denotes a total obtained by summation over that subscript. The subscript a denotes an empirical estimate and the subscript t denotes a target. The notation is consistent with that of some other Census Bureau documentation of the calculations, except that F is used in place of AF to indicate adjustment factor; some additional notation is introduced as well. In operation, Q = S = 1000.

census estimates

N_{ci} census count, area i

N_{chi} census count, part of poststratum h that is in area i

N_{c+} census count for aggregation of areas

$$N_{c+} = \sum_{i} N_{ci}$$

P_{ci} population share of area i; $P_{ci} = N_{ci}/N_{c+}$

adjusted estimates

F_{ahq} replication q of adjustment factor for poststratum h

$$\bar{F}_{ah+} = \sum_{q=1}^{Q} F_{ahq}/Q$$

F_{ah} production adjustment factor for poststratum h

$$F_{ah} = \bar{F}_{ah+}$$

$\hat{V}_{ahℓ}$ estimated covariance between F_{ah} and $F_{aℓ}$

$$\hat{V}_{ahℓ} = \sum_{q=1}^{Q} (F_{ahq} - \bar{F}_{ah+})(F_{aℓq} - \bar{F}_{aℓ+})/(Q-1)$$

X_{aiq} replication q of adjusted count for area i

$$X_{aiq} = \sum_{h} F_{ahq} N_{cih}$$

X_{a+q} replication q of adjusted count for an aggregation of areas

$$X_{a+q} = \sum_{i} X_{aiq}$$

X_{ai} production adjusted count for area i.

$$X_{ai} = \sum_{h} F_{ah} N_{cih}$$

X_{a+} adjusted count for an aggregation of areas

$$X_{a+} = \sum_{i} X_{ai}$$

P_{aiq} replication q of adjusted population share of area i; $P_{aiq} = X_{aiq}/X_{a+q}$

$\bar{P}_{ai+}$ average of replicates of adjusted share of area i; $\bar{P}_{ai+} = \sum_{q=1}^{Q} P_{aiq}/Q$

P_{ai} production adjusted population share of area i; $P_{ai} = X_{ai}/X_{a+}$

$\hat{V}_{Pai}$ estimate of variance of P_{ai}

$$\hat{V}_{Pai} = \sum_{q=1}^{Q} (P_{aiq} - \overline{P}_{ai+})^2/(Q-1)$$

targets

F_{ths} replication s of target adjustment factor for poststratum h

$$\overline{F}_{th+} = \sum_{s=1}^{S} F_{ths}/S$$

F_{th} target estimate of adjustment factor for poststratum h

$$F_{th} = \overline{F}_{th+}$$

$\hat{V}_{th\ell}$ estimated covariance between F_{th} and $F_{t\ell}$

$$\hat{V}_{th\ell} = \sum_{s}^{S} (F_{ths} - \overline{F}_{th+})(F_{t\ell s} - \overline{F}_{t\ell +})/(S-1).$$

X_{tis} replication s of target count for area i

$$X_{tis} = \sum_{h} F_{ths} N_{cih}$$

X_{t+s} replication s of target count for an aggregation of areas

$$X_{t+s} = \sum_{i} X_{tis}$$

P_{tis} replication s of target population share of area i; $P_{tis} = X_{tis}/X_{t+s}$

$\overline{P}_{ti+}$ average of replicates of target share of area i; $\overline{P}_{ti+} = \sum_{s=1}^{S} P_{tis}/S$

P_{ti} target share of area i; $P_{ti} = \bar{P}_{ti+}$

B_{Pi} estimate of bias in adjusted share, P_{ai}

$B_{Pi} = P_{ai} - P_{ti}$

$\hat{V}_{BPi}$ estimate of variance of B_{Pi}

$$\hat{V}_{BPi} = \sum_{s=1}^{S} (P_{tis} - \bar{P}_{ti+})^2/(S-1)$$

Loss Function Calculations

First consider the MSE for the census. Define

$$L_{ci} = (P_{ci} - P_{ti})^2$$

$$L_{ciqs} = [P_{ci} - P_{tis} + (P_{ai} - P_{aiq})]^2$$

$$\bar{L}_{ci++} = \sum_{s=1}^{S} \sum_{q=1}^{Q} L_{ciqs}/(QS)$$

and observe that

$$\bar{L}_{ci++} = L_{ci} + (1 - S^{-1})\hat{V}_{BPi} + (1 - Q^{-1})\hat{V}_{Pai}.$$

Thus,

$$2L_{ci} - \bar{L}_{ci+-} = L_{ci} - (1 - S^{-1})\hat{V}_{BPi} - (1 - Q^{-1})\hat{V}_{Pai},$$

as desired in (1) except for the small terms in S^{-1} and Q^{-1} (see overview), and so we estimate the MSE in the P_{ct} by $L_{ct}^{R} = 2L_{ct} - \bar{L}_{ct\cdot+}$.

Turning attention to the adjusted estimates, define

$L_{atq} = (P_{atq} - P_{t\cdot})^2$

$L_{atqs} = (P_{atq} - P_{ts})^2$

$$\bar{L}_{at\cdot} = \sum_{q=1}^{Q} L_{atq}/Q$$

$$\bar{L}_{at\cdot\cdot} = \sum_{s=1}^{S}\sum_{q=1}^{Q} L_{atqs}/(QS)$$

and observe that

$$\bar{L}_{at\cdot} = B_{Pt}^2 + (1 - Q^{-1})\hat{V}_{Pat}$$

and

$$\bar{L}_{at\cdot+} = \bar{L}_{at\cdot} + (1 - S^{-1})\hat{V}_{BPt}.$$

Thus, $2\bar{L}_{at\cdot} - \bar{L}_{at\cdot+} = B_{Pt}^2 + \hat{V}_{Pat} - \hat{V}_{BPt}$, as desired in (2), and so we estimate the MSE of P_{at} by

$L_{at}^{R} = 2\bar{L}_{at\cdot} - \bar{L}_{at\cdot\cdot}$.

Notes

Error from choice of imputation method was not reflected in $\hat{V}_{ah\ell}$. It was reflected in $\hat{V}_{BPt}$, but that does not affect the point estimates of difference in expected loss. The variance of the estimate of correlation bias is not reflected in $\hat{V}_{BPt}$.

Appendix G: Inconsistency of Poststratification between P Sample and E Sample

The classification of a person into a poststratum can be different in the census and the P-sample. This inconsistency may cause a bias in the DSE because the coverage factors for gross undercount (including not data defined persons) are derived for poststrata based on the P-sample and are applied to the poststrata based on census enumerations, or what we call E-sample poststrata. The adjustment factor for a poststratum is the product of two factors. The first factor is an adjustment for erroneous enumerations and non-data defined persons, and it involves only E-sample poststrata. The second factor is an adjustment for persons who are not enumerated (including not data defined) in the census. This factor involves only P-sample poststrata. Ideally, this factor would be based on E-sample poststratification, but of course that is not completely feasible. In this appendix we describe a method for estimating the bias from inconsistent poststratification. The method has not yet been applied.

To understand the bias, it is useful to consider that *each* person enumerated in the P- sample *could* be enumerated both ways and assigned to a poststratum two ways, based on either the P-sample data or the census data. We must make some assumptions to estimate what the E-sample poststratum is for a person enumerated in the P- sample but not the census. Specifically, we mean what E-sample poststratum would have been assigned if the person had been enumerated in the census. The reference in all cases is to Census-day characteristics, and is distinct from the differences between inmover characteristics and outmover characteristics, as the latter differences reflect the effect of change over time.

Index the E-sample and P-sample poststrata by h and k, and assume that the indexing is consistent, so that if the variables recorded for a person were consistent between the census and the P-sample, and the person were in E-sample poststratum h, the person would also be in P-sample poststratum h.

Define the following quantities.

G = a subgroup of P-sample enumerations, such as enumerations classified as inmovers, outmovers, nonmovers, persons with imputed P-sample characteristics, etc.

$f_G(h|k)$ = the proportion of group G persons enumerated in P-sample poststratum k who belong to E-sample poststratum h.

$\hat{N}_{CP,G}(h,k)$ = estimate of the number of P-sample population from group G that are in E-sample poststratum h and P-sample poststratum k and that were enumerated in the census.

$\hat{N}_{P,G}(h,k)$ = estimate of the number of P-sample population enumerations from group G that are in E-sample poststratum h and P-sample poststratum k.

The Census Bureau's estimate of the P-sample population size for group G in poststratum i is, say, $\hat{N}_{P,G}(i)$, and because it is based on P-sample poststratification it is equal to

$$\hat{N}_{P,G}(i) = \sum_h \hat{N}_{P,G}(h,i).$$

If the estimate were based on E-sample poststratification, it would be

$$\sum_k \hat{N}_{P,G}(i,k).$$

The error in $\hat{N}_{P,G}(i)$ from the inconsistent poststratification, say $n_{P,G}$, is thus

$$n_{P,G} = \hat{N}_{P,G}(i) - \sum_k \hat{N}_{P,G}(i,k).$$

The Census Bureau's estimate of the number of the P sample population in group G who were enumerated in the census in poststratum i is, say, $\hat{N}_{CP,G}(i)$, and because it is based on P-sample poststratification it is equal to

$$\hat{N}_{CP,G}(i) = \sum_h \hat{N}_{CP,G}(h,i).$$

If the estimate were based on E-sample poststratification, it would be

$$\sum_k \hat{N}_{CP,G}(i,k).$$

The error in $\hat{N}_{CP,G}(i)$ from the inconsistent poststratification, say $n_{CP,G}(1)$, is thus

$$n_{CP,G}(i) = \hat{N}_{CP,G}(i) - \sum_k \hat{N}_{CP,G}(i,k).$$

To estimate the means of $n_{P,G}$ and $n_{CP,G}$ we could use estimates of f_G, say $\hat{f}_G$. These estimates are being developed but even rough estimates will not be available until after March 1. Once they are available, we would estimate the expected value of $n_{P,G}(i)$ by

$$\hat{n}_{P,G}(i) = \hat{N}_{P,G}(i) - \sum_k \hat{f}_G(i|k)\hat{N}_{P,G}(k)$$

and estimate the expected value of $n_{CP,G}(i)$ by

$$\hat{n}_{CP,G}(i) = \hat{N}_{CP,G}(i) - \sum_k \hat{f}_G(i|k)\hat{N}_{CP,G}(k).$$

Alternative estimates of $\hat{N}_{P,G}(i)$ and $\hat{N}_{CP,G}(1)$ could then be obtained as $\hat{N}_{P,G}(i) - \hat{n}_{P,G}(i)$ and $\hat{N}_{CP,G}(i) - \hat{n}_{CP,G}(i)$. These alternative estimates could be used to recalculate adjustment factors, and the

effect on estimates of population size and population shares could be analyzed.

Appendix H: Confidence Intervals

We constructed confidence intervals for the net undercount rate in such a way as to allow for the estimated net bias in the estimates of undercount. The estimates of the net bias and a component of the covariance matrix of the DSEs and the variance of the nonsampling bias are based on simulations with 1000 replications for the 416 A.C.E. poststrata (see Appendix F for details). Originally there were 448 poststrata, but some collapsing was done for variance reduction. The estimation of the total variance V is described in Appendix C. We estimate the net bias in the DSE by the difference between the DSE observed in the A.C.E. and the mean of the replicated values. The net bias in the net undercount rate, $B(\hat{U})$, is estimated similarly. With the estimated bias and variance, we form the 95% confidence interval for the net undercount rate by

$$(\hat{U} - \hat{B}(\hat{U}) - 2V^{1/2}, \hat{U} - \hat{B}(\hat{U}) + 2V^{1/2}).$$

Tables 20 through 23 and Figures 1 through 4 show the 95% confidence intervals for the 16 evaluation poststrata for the four set of assumptions about correlation bias discussed in Appendix D. These tables and figures contain confidence intervals when the bias is distributed from the intermediate poststrata to the A.C.E. poststrata proportional to the DSE, called GROSDSE. The confidence intervals obtained when the bias is distributed proportional to the gross undercount, called GROSUC, are practically the same and are not shown.

The 95% confidence intervals for the undercount rate shown in Figure 3 and Table 22 reflect our preferred set of component errors. The confidence intervals for eight of the 16 evaluation poststrata cover zero. For Evaluation Poststrata 3 and 7, the confidence intervals do not cover the estimated undercount rate from the A.C.E. In both cases, the confidence intervals imply that the A.C.E. is overestimating the undercount rate. Since the estimate shows an overcount for Evaluation Poststratum 3, the confidence interval indicates a larger overcount. The confidence interval for the

undercount rate for the U.S. does not cover zero and A.C.E. estimate is on the upper edge of the interval.

The confidence intervals indicate an undercount does exist for minorities and for renters. The supporting evidence is that five of the six minority evaluation poststrata (No.11-16) do not cover zero, and five of the six non-owner poststrata (No.8-10, 14-16) do not cover zero. As a sensitivity analysis, even under the assumption of no correlation bias, the undercount for renters appears to exist with the same evaluation poststrata covers zero and an additional minority renter evaluation poststrata just touches zero. However, with the assumption of no correlation bias, the three owner minority poststrata cover zero. The confidence interval for the undercount rate for the U.S. still does not cover zero with the assumption of no correlation bias.

We also have used the simulation methodology to examine the individual effect of the sampling and nonsampling errors on the undercount rate at the national level, assuming all other errors are zero. Table 24 shows the bias, standard error, and root mean square error ($MSE^{1/2}$) for each component error. Using the root mean square error to rank error sources, the major contributors to bias are P-sample collection error and correlation bias. Following are E-sample collection error, E-sample processing error and P-sample matching error. Sampling error, P-sample fabrication error, ratio estimator bias, and imputation error are lower still. Correlation bias and E-sample collection error introduce a negative bias, causing the DSE to be an underestimate of the population size while the other components introduce a positive bias.

In Tables 3 through 18, $\hat{B}(\hat{U})$ describes the individual effect of the error components on the bias of the net undercount rate at the evaluation poststratum level. The bias estimate $\hat{B}(\hat{U})$ shown for each error component is calculated algebraically, not by simulation, under the assumption that only one source of error is present. The contribution of the individual error components to the nonsampling variance at the evaluation poststratum level is not shown.

Table 1. 16 Evaluation Poststrata			
		No. in MVF P-sample (1990)	PS Groups (2000)
1. Non-minority/owner/large and Medium MSA MO-MB NE/MW	high RR	4,960	1,2,9,10
2. Non-minority/owner/large and Medium MSA MO-MB S/W	high RR	7,702	3,4,11,12
3. Non-minority/owner/large and Medium MSA MO-MB NE/MW	low RR	3,031	5,6,13,14
4. Non-minority/owner/large and Medium MSA MO-MB S/W	low RR	2,936	7,8,15,16
5. Non-minority/owner/Small MSA and Non-MSA MO-MB	high RR	5,560	17-20
6 Non-minority/owner/ Small MSA and Non-MSA MO-MB	low RR	2,095	21-24
7. Non-minority/Owner/All Other TEAs		7,355	25-32
8. Non-minority/ Non-Owner/Large or Medium MSA MO-MB	high RR	4,963	33, 35
9.Non-minority/ Non-Owner/Large or Medium MSA MO-MB	low RR	3,197	34, 36
10. Non-minority/non-owner/Small MSA & Non-MSA MO-MB All other TEA		5,291	37-40
11. Minority/owner/large and Medium MSA MO-MB	high RR	8,841	41, 49, 57,59
12. Minority/owner/large and Medium MSA MO-MB	low RR	5,628	42, 50
13. Minority/Owner/All Other TEAs		3,877	43, 44, 51, 52
14. Minority/ Non-Owner/Large or Medium MSA MO-MB	high RR	10,809	45, 53, 58,60
15. Minority/ Non-Owner/Large or Medium MSA MO-MB	low RR	6,421	46, 54
16. Minority/Non-Owner/All Other TEAs		3,797	47, 48, 55, 56, 61-64
Total		86,463	

Table 2. Sources of Data for Estimation of Components of Error

Error Components	Measurement in 1990	Measurement in 2000
P-sample matching error	1990 Matching Error Study	1990 Matching Error Study with adjustments for 2000
P-sample data collection error	1990 Evaluation Followup	1990 Evaluation Followup with adjustments for 2000
P-sample fabrication	1990 Evaluation Followup	1990 Evaluation Followup with adjustments for 2000
E-sample data collection error	1990 Evaluation Followup	1990 Evaluation Followup with adjustments for 2000
E-sample processing error	1990 Matching Error Study	1990 Matching Error Study with adjustments for 2000
Correlation bias	1990 Demographic Analysis	2000 Demographic Analysis
Ratio estimator bias	1990 PES	2000 A.C.E.
Sampling error	1990 PES	2000 A.C.E.
Imputation error	1990 Reasonable Alternatives Imputation Study	1990 Reasonable Alternatives with adjustments for 2000
Excluded Census Data Error	1990 Excluded Data Study	Not available
Contamination of P sample by enumeration or vice versa	Shown to be negligible	Not available in time for analysis for decision
Misclassification error of records into poststrata from inconsistent reporting	Not measured	Not available in time for analysis for decision
Synthetic error	Artificial population analysis and not integrated in total error model	Under development but will not be integrated in total error model

Table 3. 1990 Undercount Rates for 2000 Evaluation Poststrata

2000 Evaluation Poststrata	2000 UC	1990 UC
US	1.1788	1.7500
1.N-min/own/lrg&med MSA/MO-MB-hi/NE/MW	0.2695	-1.0324
2 N-min/own/lrg&med MSA/MO-MB-hi/S/W	0.0947	-0 2473
3.N-min/own/lrg&med MSA/MO-MB-lo/NE/MW	-2.8191	-0.2034
4.N-min/own/lrg&med MSA/MO-MB-lo/S/W	1.2840	1.1349
5.N-min/own/small MSA/MO-MB-hi	0.2127	-1.0053
6.N-min/own/small MSA/MO-MB-lo	2.3302	0.3729
7.N-min/own/All other TEAs	0.4232	0.4382
8.N-min/n-own/lrg&med MSA/MO-MB-hi	1.1290	2.9432
9.N-min/n-own/lrg&med MSA/MO-MB-lo	1.8404	4.5272
10.N-min/n-own/small MSA/MO-MB&other TEAs	2.5867	3.6684
11.Min/own/lrg&med MSA/MO-MB-hi	1.3307	1.1367
12.Min/own/lrg&med MSA/MO-MB-lo	-0.6778	2.3954
13.Min/own/All other TEAs	0.7719	2.5138
14.Min/n-own/lrg&med MSA/MO-MB-hi	3.5018	7.9802
15.Min/n-own/lrg&med MSA/MO-MB-lo	4.2140	6 5061
16.Min/n-own/All other TEAs	3.9699	6.0418

Table 4.
Moments of Error Components for
Evaluation Poststratum 01
Non-min/owner/Large or Medium
MSA - High - NE/MW

Nce = 34,282,550 Direct DSE = 35,646,814

Np = 35,984,939 $\hat{U}$ = 0.271

Ncp = 34,607,734 Census = 35,550,177

Error Source	Pos Gross Error	Neg Gross Error	Net Error	$\hat{B}(\hat{U})$
Matching Error				
Mm	79,467	160,433	(80,966)	0.29
	(46,175)	(49,853)	(58,746)	
Npm	71,576	49,913	21663	
	(27,714)	(25,519)	(10,508)	
P Sample Collection				
Npa	297,615	38,881	258,734	0.28
	(232,050)	(16,135)	(231,518)	
Ma	242,628	89,648	152,980	
	(229,502)	(39,140)	(234,747)	
P Sample Fabrication				
Npf		0	0	0.00
		0	0	
Mf		0	0	
		0	0	
E Sample Error				
Co	118,562	104,287	14,275	0.04
	(67,514)	(49,188)	(83,466)	
Cc	50,031	128,833	(78,803)	-0.23
	(19,471)	(67,035)	(70,132)	
Model Bias (Tau)				
Imputation Error				
Ratio Estimator Bias			4,608	0.01
			461	
Net Sampling Error			0	
			(80,076)	

Table 5.
Moments of Error Components for
Evaluation Poststratum 02
Non-min/owner/Large or Medium
MSA - High - S/W

Nce = 29,785,508 Direct DSE = 31,284,669

Np = 32,448,004 $\hat{U}$ = 0.102

Ncp = 30,893,095 Census = 31,252,841

Error Source	Pos Gross Error	Neg Gross Error	Net Error	$\hat{B}(\hat{U})$
Matching Error				
Mm	65,321	95,909	-30588	0.01
	(32,990)	(33,361)	(46,830)	
Npm	35,732	64,360	-28628	
	(15,551)	(33,111)	(36,549)	
P Sample Collection				
Npa	59,015	36,486	22,529	0.33
	(29,549)	(17,011)	(18,124)	
Ma	41,316	120,654	-79339	
	(41,316)	(66,891)	(38,580)	
P Sample Fabrication				
Npf		5,168	5,168	0.04
		(5,168)	(5,168)	
Mf		17,377	17,377	
		(17,377)	(17,377)	
E Sample Error				
Co	55,759	36,447	19,312	0.06
	(36,038)	(13,756)	(38,438)	
Cc	34,409	65,455	-31,046	-0.10
	(25,522)	(31,415)	(40,787)	
Model Bias (Tau)				
Imputation Error				
Ratio Estimator Bias			553	0.01
			55	
Net Sampling Error			0	
			(79,854)	

Table 6.
Moments of Error Components for
Evaluation Poststratum 03
Non-min/owner/Large or Medium
MSA - Low - NE/MW

Nce = 4,893,801 Direct DSE = 5,217,719

Np = 4,811,175 $\hat{U}$ = -2.941

Ncp = 4,512,495 Census = 5,371,168

Error Source	Pos Gross Error	Neg Gross Error	Net Error	$\hat{B}(\hat{U})$
Matching Error				
Mm	6,283	14,788	-8,504	0.31
	(4,191)	(6,554)	(5,396)	
Npm	7,770	2,366	5,404	
	(5,290)	(1,579)	(5,545)	
P Sample Collection				
Npa	9,723	8,197	1,526	0.75
	(5,992)	(5,605)	(7,203)	
Ma	379	31,983	-31,604	
	(379)	(19,963)	(19,967)	
P Sample Fabrication				
Npf		0	0	0.00
		(0)	(0)	
Mf		0	0	
		(0)	(0)	
E Sample Error				
Co	71,698	13,777	57,921	1.23
	(49,900)	(6,220)	(50,269)	
Cc	29,353	23,113	6,239	0.13
	(15,591)	(13,733)	(17,748)	
Model Bias (Tau)				
Imputation Error				
Ratio Estimator Bias			120	0.00
			(12)	
Net Sampling Error			0	
			(33,391)	

Table 7.
Moments of Error Components for
Evaluation Poststratum 04
Non-min/owner/Large or Medium
MSA -Low - S/W

Nce = 7,716,712 Direct DSE = 8,358,892

Np = 7,173,516 $\hat{U}$ = 1.275

Ncp = 6,622,403 Census = 8,252,306

Error Source	Pos Gross Error	Neg Gross Error	Net Error	$\hat{B}(\hat{U})$
Matching Error				
Mm	28,951	19,464	9,487	-0.08
	(13,219)	(9,074)	(16,333)	
Npm	23,231	18,491	4,741	
	(13,969)	(9,105)	(13,041)	
P Sample Collection				
Npa	123,199	122,715	484	0.06
	(111,652)	(107,867)	(10,659)	
Ma	106,748	110,215	-3,468	
	(105,819)	(109,025)	(51,235	
P Sample Fabrication				
Npf		0	0	0.00
		(0)	(0)	
Mf		0	0	
		(0)	(0)	
E Sample Error				
Co	45,454	47,054	-1,600	-0.02
	(19,432)	(25,323)	(32,585)	
Cc	2,216	53,934	-51,718	-0.66
	(1,989)	(39,686)	(39,736)	
Model Bias (Tau)				
Imputation Error				
Ratio Estimator Bias			1,113	0.01
			(111)	
Net Sampling Error			0	
			(83,149)	

Table 8.
Moments of Error Components for Evaluation Poststratum 05 Non-min/owner/Small and Non-MSA - High

Nce = 24,649,632 Direct DSE = 25,751,566

Np = 25,377,448 $\hat{U}$ = 0.209

Ncp = 24,291,523 Census = 25,697,696

Error Source	Pos Gross Error	Neg Gross Error	Net Error	$\hat{B}(\hat{U})$
Matching Error				
Mm	31,555	63,078	-31,523	0.17
	(13,256)	(21,561)	(24,222)	
Npm	29,184	18,333	10,851	
	(11,944)	(8,293)	(14,661	
P Sample Collection				
Npa	29,131	20,600	8,531	0.08
	(13,518)	(13,855)	(12,258)	
Ma	6,513	16,627	-10,114	
	(6,513)	(9,888)	(11,839)	
P Sample Fabrication				
Npf		0	0	0.00
		(0)	(0)	
Mf		0	0	
		(0)	(0)	
E Sample Error				
Co	94,940	33,672	61,269	0.25
	(35,208)	(18,906)	(37,892)	
Cc	21,980	43,677	-21,697	-0.09
	(11,606)	(18,101)	(21,518)	
Model Bias (Tau)				
Imputation Error				
Ratio Estimator Bias			11,291	0.04
			(1,129)	
Net Sampling Error			0	
			(72,062)	

Table 9.
Moments of Error Components for Evaluation Poststratum 06 Non-min/owner/Small and Non-MSA - Low

Nce = 5,817,573 Direct DSE = 6,338,959

Np = 6,441,327 $\hat{U}$ = 2.203

Ncp = 5,911,522 Census = 6,199,286

Error Source	Pos Gross Error	Neg Gross Error	Net Error	$\hat{B}(\hat{U})$
Matching Error				
Mm	16,208	21,084	-4,876	0.06
	(9,084)	(12,779)	(14,539)	
Npm	14,590	16,007	-1,417	
	(9,597)	(8,857)	(7,006)	
P Sample Collection				
Npa	24,405	17,576	6,828	0.36
	(16,086)	(10,715)	(6,828)	
Ma	8,856	24,542	-15,686	
	(6,418)	(13,039)	(10,292)	
P Sample Fabrication				
Npf		0	0	0.00
		(0)	(0)	
Mf		0	0	
		(0)	(0)	
E Sample Error				
Co	61,037	45,235	15,802	0.27
	(34,009)	(22,425)	(17,298)	
Cc	16,105	33,286	-17,180	-0.29
	(15,250)	(21,584)	(9,343)	
Model Bias (Tau)				
Imputation Error				
Ratio Estimator Bias			415	0.01
			(42)	
Net Sampling Error			0	
			(51,526)	

Table 10.
Moments of Error Components for
Evaluation Poststratum 07
Non-min/owner/All Other TEAs

Nce = 32,195,096 Direct DSE = 34,773,055

Np = 32,656,527 $\hat{U}$ = 0.401

Ncp = 30,235,481 Census = 34,633,612

Error Source	Pos Gross Error	Neg Gross Error	Net Error	$\hat{B}(\hat{U})$
Matching Error				
Mm	85,548	168,896	-83,349	0.23
	(40,951)	(46,856)	(38,564)	
Npm	83,393	96,893	-13,501	
	(31,865)	(43,863)	(21,688)	
P Sample Collection				
Npa	160,135	202,065	-41,930	0.79
	(44,847)	(82,204)	(73,861)	
Ma	42,081	320,702	-278,621	
	(15,982)	(132,414)	(131,817)	
P Sample Fabrication				
Npf		0	0	0.00
		(0)	(0)	
Mf		0	0	
		(0)	(0)	
E Sample Error				
Co	269,788	114,722	155,067	0.48
	(125,846)	(42,075)	(109,567)	
Cc	24,602	100,381	-75,778	-0.23
	(15,546)	(39,392)	(34,993)	
Model Bias (Tau)				
Imputation Error				
Ratio Estimator Bias			13,366	0.04
			(1,337)	
Net Sampling Error			0	
			(124,413)	

Table 11.
Moments of Error Components for
Evaluation Poststratum 08
Non-min/non-owner/Large or
Medium MSA - High

Nce = 18,112,506 Direct DSE = 20,213,083

Np = 19,175,297 $\hat{U}$ = 1.097

Ncp = 17,182,568 Census = 19,991,324

Error Source	Pos Gross Error	Neg Gross Error	Net Error	$\hat{B}(\hat{U})$
Matching Error				
Mm	34,026	26,913	7,113	0.01
	(14,091)	(15,149)	(17,050)	
Npm	29,384	20,120	9,264	
	(13,256)	(10,935)	(10,535)	
P Sample Collection				
Npa	63,198	93,825	-30,627	0.38
	(23,359)	(40,768)	(29,128)	
Ma	24,341	117,647	-93,305	
	(13,343)	(69,209)	(57,457)	
P Sample Fabrication				
Npf		0	0	0.00
		(0)	(0)	
Mf		0	0	
		(0)	(0)	
E Sample Error				
Co	42,893	44,076	-1,183	-0.01
	(16,008)	(24,311)	(28,123)	
Cc	47,640	105,936	-58,296	-0.32
	(19,356)	(41,402)	(46,174)	
Model Bias (Tau)				
Imputation Error				
Ratio Estimator Bias			-1,264	-0.01
			126	
Net Sampling Error			0	
			(100,570)	

Table 12.
Moments of Error Components for
Evaluation Poststratum 09
Non-min/non-owner/Large or
Medium MSA - Low

Nce = 6,023,062 Direct DSE = 7,035,171

Np = 6,468,268 $\hat{U}$ = 1.799

Ncp = 5,537,716 Census = 6,908,574

Error Source	Pos Gross Error	Neg Gross Error	Net Error	$\hat{B}(\hat{U})$
Matching Error				
Mm	19,705	41,141	-21,436	0.50
	(7,604)	(14,417)	(15,543)	
Npm	24,135	16,092	8,044	
	(12,088)	(6,618)	(13,428)	
P Sample Collection				
Npa	19,360	39,218	-19,858	0.16
	(9,346)	(33,993)	(34,962)	
Ma	8,439	34,344	-25,905	
	(5,976)	(32,107)	(32,582)	
P Sample Fabrication				
Npf		0	0	0.00
		(0)	(0)	
Mf		0	0	
		(0)	(0)	
E Sample Error				
Co	62,054	33,753	28,300	0.46
	(22,099)	(12,581)	(21,750)	
Cc	18,456	49,762	-31,306	-0.51
	(8,243)	(20,936)	(22,541)	
Model Bias (Tau)				
Imputation Error				
Ratio Estimator Bias			322	0.00
			(32)	
Net Sampling Error			0	
			(56,793)	

Table 13.
Moments of Error Components for
Evaluation Poststratum 10
Non-min/non-owner/Small and
Non-MSA, All Other TEAs

Nce = 17,212,267 Direct DSE = 19,551,600

Np = 18,265,774 $\hat{U}$ = 2.479

Ncp = 16,080,289 Census = 19,067,004

Error Source	Pos Gross Error	Neg Gross Error	Net Error	$\hat{B}(\hat{U})$
Matching Error				
Mm	51,055	206,208	-155,153	1.07
	(16,751)	(80,344)	(79,702)	
Npm	67,768	44,465	23,304	
	(27,222)	(17,623)	(23,523)	
P Sample Collection				
Npa	84,539	98,049	-13,510	0.62
	(25,718)	(33,980)	(27,515)	
Ma	24,531	138,758	-114,227	
	(11,665)	(42,276)	(39,571)	
P Sample Fabrication				
Npf		0	0	0.00
		(0)	(0)	
Mf		0	0	
		(0)	(0)	
E Sample Error				
Co	196,274	93,909	102,366	0.58
	(79,354)	(32,660)	(85,676)	
Cc	136,021	351,845	-215,824	-1.21
	(50,836)	(190,221)	(196,523)	
Model Bias (Tau)				
Imputation Error				
Ratio Estimator Bias			4,381	0.02
			(438)	
Net Sampling Error			0	
			(88,933)	

Table 14.
Moments of Error Components for
Evaluation Poststratum 11
Minority/owner/Large or Medium
MSA - High

Nce = 22,815,631 Direct DSE = 24,896,228

Np = 23,316,868 $\hat{U}$ = 1.284

Ncp = 21,368,260 Census = 24,576,535

Error Source	Pos Gross Error	Neg Gross Error	Net Error	$\hat{B}(\hat{U})$
Matching Error				
Mm	32,919	69,315	-36,396	0.10
	(10,981)	(23,217)	(25,823)	
Npm	22,469	37,631	-15,162	
	(10,278)	(10,799)	(14,812)	
P Sample Collection				
Npa	228,619	34,430	194,189	0.67
	(100,750)	(15,467)	(99,953)	
Ma	103,727	70,384	33,343	
	(92,085)	(30,703)	(95,865)	
P Sample Fabrication				
Npf		292	-292	0.00
		(292)	(292)	
Mf		269	-269	
		(269)	(269	
E Sample Error				
Co	74,568	16,690	57,878	0.25
	(38,564)	(6,536)	(39,074)	
Cc	39,425	46,790	-7,365	-0.03
	(11,436)	(21,650)	(23,803)	
Model Bias (Tau)				
Imputation Error				
Ratio Estimator Bias			8,683	0.03
			(868)	
Net Sampling Error			0	
			(98,342)	

Table 15.
Moments of Error Components for
Evaluation Poststratum 12
Minority/owner/Large or Medium
MSA - Low

Nce = 4,620,389 Direct DSE = 5,285,962

Np = 4,532,239 $\hat{U}$ = -0.765

Ncp = 3,961,569 Census = 5,326,380

Error Source	Pos Gross Error	Neg Gross Error	Net Error	$\hat{B}(\hat{U})$
Matching Error				
Mm	9,165	22,438	-13,273	0.61
	(6,192)	(9,053)	(10,999)	
Npm	19,938	7,767	12,171	
	(9,556)	(6,086)	(11,328)	
P Sample Collection				
Npa	9,666	51,283	-41,617	0.52
	(3,946)	(40,001)	(40,202)	
Ma	406	57,469	-57,063	
	(406)	(43,042)	(43,044)	
P Sample Fabrication				
Npf		0	0	0.00
		(0)	(0)	
		0	0	
		(0)	(0)	
E Sample Error				
Co	16,157	14,033	2,124	0.05
	(4,890)	(3,901)	(4,955)	
Cc	14,196	19,248	-5,053	-0.11
	(4,226)	(6,144)	(6,927)	
Model Bias (Tau)				
Imputation Error				
Ratio Estimator Bias			296	0.01
			(30)	
Net Sampling Error			0	
			(45,411)	

Table 16.
Moments of Error Components for
Evaluation Poststratum 13
Minority/owner/All Other TEAs

Nce = 8,859,679 Direct DSE = 9,841,047

Np = 8,697,210 $\hat{U}$ = 0.651

Ncp = 7,829,907 Census = 9,776,940

Error Source	Pos Gross Error	Neg Gross Error	Net Error	$\hat{B}(\hat{U})$
Matching Error				
Mm	18,377	13,996	4,382	-0.19
	(8,636)	(7,431)	(11,354)	
Npm	10,642	22,324	-11,681	
	(5,917)	(11,875)	(13,130	
P Sample Collection				
Npa	183,215	130,607	52,607	0.43
	(133,248)	(130,607)	(27,946)	
Ma	22,904	9,264	13,640	
	(22,904)	(6,598)	(23,887)	
P Sample Fabrication				
Npf		0	0	0.00
		(0)	(0)	
Mf		0	0	
		(0)	(0)	
E Sample Error				
Co	31,697	21,524	10,173	0.11
	(13,204)	(10,925)	(17,258)	
Cc	78,809	14,080	64,728	0.73
	(31,658)	(7,584)	(33,160)	
Model Bias (Tau)				
Imputation Error				
Ratio Estimator Bias			1,140	0.01
			(114)	
Net Sampling Error			0	
			(68,949)	

Table 17.
Moments of Error Components for
Evaluation Poststratum 14
Minority/non-owner/Large or
Medium MSA - High

Nce = 21,443,656 Direct DSE = 24,992,574

Np = 21,403,543 $\hat{U}$ = 3.341

Ncp = 18,364,263 Census = 24,157,485

Error Source	Pos Gross Error	Neg Gross Error	Net Error	$\hat{B}(\hat{U})$
Matching Error				
Mm	67,274	121,786	-54,511	0.34
	(23,216)	(25,173)	(32,640)	
Npm	72,090	60,142	11,948	
	(25,032)	(23,639)	(34,478)	
P Sample Collection				
Npa	91,449	82,514	8,934	-0.02
	(37,729)	(49,987)	(60,023)	
Ma	108,725	97,066	11,659	
	(74,071)	(51,135)	(91,590)	
P Sample Fabrication				
Npf		20,912	-20,912	0.22
		(20,912)	(20,912)	
Mf		60,413	-60,413	
		(60,413)	(60,413)	
E Sample Error				
Co	57,648	76,277	-18,629	-0.08
	(19,758)	(23,852)	(29,551)	
Cc	68,998	96,545	-27,547	-0.12
	(25,177)	(23,469)	(34,135)	
Model Bias (Tau)				
Imputation Error				
Ratio Estimator Bias			389	0.00
			(39)	
Net Sampling Error			0	
			(119,134)	

Table 18.
Moments of Error Components for
Evaluation Poststratum 15
Minority/non-owner/Large or
Medium MSA - Low

Nce = 6,310,050 Direct DSE = 7,803,395

Np = 7,660,305 $\hat{U}$ = 4.052

Ncp = 6,194,343 Census = 7,487,171

Error Source	Pos Gross Error	Neg Gross Error	Net Error	$\hat{B}(\hat{U})$
Matching Error				
Mm	25,312	77,351	-52,039	0.81
	(6,795)	(30,179)	(27,662)	
Npm	20,110	19,728	382	
	(6,474)	(6,029)	(8,303)	
P Sample Collection				
Npa	29,114	22,751	6,363	0.33
	(15,940)	(9,852)	(14,494)	
Ma	13,095	28,921	-15,826	
	(9,066)	(10,768)	(8,021	
P Sample Fabrication				
Npf		1,172	-1,172	0.03
		(919)	(919)	
Mf		2,884	-2,884	
		(2,211)	(2,211)	
E Sample Error				
Co	60,172	41,682	18,491	0.28
	(27,583)	(15,084)	(24,118)	
Cc	36,168	42,828	-6,659	-0.10
	(15,700)	(12,730)	(15,593)	
Model Bias (Tau)				
Imputation Error				
Ratio Estimator Bias			1,506	0.02
			(151)	
Net Sampling Error			0	
			(65,577)	

Table 19.
Moments of Error Components for
Evaluation Poststratum 16
Minority/non-owner/All Other
TEAs

Nce = 8,229,779 Direct DSE = 9,718,222

Np = 8,386,177 $\hat{U}$ = 3.907

Ncp = 7,101,750 Census = 9,338,498

Error Source	Pos Gross Error	Neg Gross Error	Net Error	$\hat{B}(\hat{U})$
Matching Error				
Mm	46,200	78,906	-32,706	0.23
	(20,419)	(32,231)	(37,871)	
Npm	38,015	56,210	-18,195	
	(30,391)	(32,901)	(15,061)	
P Sample Collection				
Npa	42,315	14,861	27,454	1.55
	(14,485)	(13,068)	(19,479)	
Ma	11,339	102,076	-90,737	
	(7,899)	(46,791)	(47,398)	
P Sample Fabrication				
Npf		9,366	-9,366	0.00
		(9,366)	(9,366)	
Mf		8,070	-8,070	
		(8,070)	(8,070)	
E Sample Error				
Co	74,094	47,069	27,024	0.32
	(27,839)	(24,685)	(37,285)	
Cc	51,700	49,975	1,726	0.02
	(24,020)	(32,581)	(36,183)	
Model Bias (Tau)				
Imputation Error				
Ratio Estimator Bias			1,235	0.01
			(124)	
Net Sampling Error			0	
			(74,985)	

Table 20. Total error for net undercount rate assuming no correlation bias

Ev post	prod uc	corr uc	bias(uc)	se(bias)	se(pr uc)	total se 95% conf interval
US	1.1788	0.464	0.7148	0.086	0.1349	0.16 (0.1439, 0.7840)
1	0.2695	-0.1217	0.3912	0.2181	0.224	0.3127 (-0.7470, 0.5036)
2	0.0947	-0.2516	0.3463	0.1802	0.255	0.3123 (-0.8761, 0.3729)
3	-2.8191	-5.2887	2.4696	0.4999	0.6572	0.8257 (-6.9401, -3.6373)
4	1.284	1.9862	-0.7022	0.6324	0.9813	1.1674 (-0.3486, 4.3209)
5	0.2127	-0.207	0.4197	0.1047	0.2792	0.2982 (-0.8034, 0.3894)
6	2.3302	1.8476	0.4826	0.2876	0.793	0.8436 (0.1605, 3.5347)
7	0.4232	-0.853	1.2762	0.2266	0.3562	0.4222 (-1.6973, -0.0087)
8	1.129	1.0745	0.0545	0.1754	0.4918	0.5221 (0.0303, 2.1187)
9	1.8404	1.2102	0.6302	0.3538	0.7921	0.8675 (-0.5248, 2.9453)
10	2.5867	1.5337	1.053	0.54	0.4426	0.6982 (0.1373, 2.9302)
11	1.3307	0.3131	1.0177	0.2522	0.3897	0.4642 (-0.6153, 1.2414)
12	-0.6778	-1.7953	1.1176	0.4734	0.8642	0.9853 (-3.7660, 0.1753)
13	0.7719	-0.3231	1.095	0.4806	0.6944	0.8445 (-2.0120, 1.3659)
14	3.5018	3.1517	0.3502	0.386	0.4592	0.5999 (1.9519, 4.3515)
15	4.214	2.8633	1.3507	0.4191	0.8036	0.9063 (1.0506, 4.6760)
16	3.9699	1.7715	2.1984	0.4931	0.7404	0.8895 (-0.0075, 3.5505)

Table 21. Total error of net undercount rate assuming no correlation bias for Nonblack males

Ev post	prod uc	corr uc	bias(uc)	se(bias)	se(prod uc)	total se	95% Conf Interval
US	1.1788	0.7338	0.445	0.0859	0.1349	0.1599	(0.4140, 1.0536)
1	0.2695	-0.1217	0.3912	0.2181	0.224	0.3127	(-0.7470, 0.5036)
2	0.0947	-0.2516	0.3463	0.1802	0.255	0.3123	(-0.8761, 0.3729)
3	-2.8191	-5.2887	2.4696	0.4999	0.6572	0.8257	(-6.9401, -3.6373)
4	1.284	1.9862	-0.7022	0.6324	0.9813	1.1674	(-0.3486, 4.3209)
5	0.2127	-0.207	0.4197	0.1047	0.2792	0.2982	(-0.8034, 0.3894)
6	2.3302	1.8476	0.4826	0.2876	0.793	0.8436	(0.1605, 3.5347)
7	0.4232	-0.853	1.2762	0.2266	0.3562	0.4222	(-1.6973, -0.0087)
8	1.129	1.0745	0.0545	0.1754	0.4918	0.5221	(0.0303, 2.1187)
9	1.8404	1.2102	0.6302	0.3538	0.7921	0.8675	(-0.5248, 2.9453)
10	2.5867	1.5337	1.053	0.54	0.4426	0.6982	(0.1373, 2.9302)
11	1.3307	1.1197	0.2111	0.25	0.3897	0.463	(0.1936, 2.0457)
12	-0.6778	-0.5303	-0.1475	0.4678	0.8642	0.9826	(-2.4955, 1.4350)
13	0.7719	0.9019	-0.13	0.4796	0.6944	0.8439	(-0.7858, 2.5897)
14	3.5018	3.935	-0.4331	0.3823	0.4592	0.5975	(2.7399, 5.1300)
15	4.214	3.8135	0.4006	0.4148	0.8036	0.9044	(2.0048, 5.6222)
16	3.9699	2.5939	1.376	0.489	0.7404	0.8873	(0.8193, 4.3684)

Table 22. Total error of net undercount rate if no correlation bias for 18-29 Nonblack males

Ev post	prod uc	corr uc	bias(uc)	se(bias)	se(prod uc)	total se 95% conf interval
US	1.1788	0.8567	0.3221	0.0857	0.1349	0.1598 (0.5370, 1.1763)
1	0.2695	0.0413	0.2282	0.2176	0.224	0.3123 (-0.5834, 0.6660)
2	0.0947	-0.0766	0.1714	0.18	0.255	0.3121 (-0.7009, 0.5477)
3	-2.8191	-5.1012	2.2821	0.499	0.6572	0.8252 (-6.7516, -3.4508)
4	1.284	2.1584	-0.8745	0.6308	0.9813	1.1665 (-0.1746, 4.4915)
5	0.2127	-0.0282	0.2409	0.1046	0.2792	0.2982 (-0.6245, 0.5681)
6	2.3302	2.0222	0.308	0.2867	0.793	0.8432 (0.3357, 3.7086)
7	0.4232	-0.6727	1.0958	0.226	0.3562	0.4219 (-1.5164, 0.1710)
8	1.129	1.1742	-0.0452	0.175	0.4918	0.522 (0.1302, 2.2182)
9	1.8404	1.3175	0.523	0.3535	0.7921	0.8674 (-0.4173, 3.0522)
10	2.5867	1.6215	0.9652	0.5396	0.4426	0.6979 (0.2257, 3.0173)
11	1.3307	1.1906	0.1402	0.2497	0.3897	0.4628 (0.2649, 2.1163)
12	-0.6778	-0.4871	-0.1907	0.4674	0.8642	0.9824 (-2.4519, 1.4778)
13	0.7719	0.9453	-0.1734	0.4792	0.6944	0.8437 (-0.7421, 2.6327)
14	3.5018	3.9754	-0.4736	0.3819	0.4592	0.5973 (2.7808, 5.1699)
15	4.214	3.843	0.371	0.4147	0.8036	0.9043 (2.0345, 5.6516)
16	3.9699	2.6335	1.3364	0.4888	0.7404	0.8871 (0.8592, 4.4078)

Table 23. Total error for net undercount rate assuming correlation bias of 2% overcount for 18-29 NB males							
Ev post	prod uc	corr uc	bias(uc)	se(bias)	se(prod uc)	total se	95% conf interval
US	1.1788	0.721	0.4578	0.0857	0.1349	0.1598	(0.4013, 1.0407)
1	0.2695	-0.0548	0.3243	0.2179	0.224	0.3125	(-0.6798, 0.5702)
2	0.0947	-0.1638	0.2585	0.1801	0.255	0.3122	(-0.7882, 0.4606)
3	-2.8191	-5.2214	2.4023	0.4996	0.6572	0.8256	(-6.8725, -3.5703)
4	1.284	2.0528	-0.7689	0.6313	0.9813	1.1668	(0.2808, 4.3865)
5	0.2127	-0.1252	0.3379	0.1047	0.2792	0.2982	(-0.7215, 0.4712)
6	2.3302	1.9149	0.4153	0.2869	0.793	0.8433	(0.2283, 3.6015)
7	0.4232	-0.7775	1.2007	0.2262	0.3562	0.422	(-1.6214, 0.0664)
8	1.129	0.929	0.2	0.1757	0.4918	0.5222	(-0.1155, 1.9734)
9	1.8404	1.0214	0.819	0.3555	0.7921	0.8682	(-0.7150, 2.7578)
10	2.5867	1.3541	1.2326	0.541	0.4426	0.699	(-0.0438, 2.7520)
11	1.3307	1.0932	0.2376	0.2494	0.3897	0.4627	(0.1679, 2.0185)
12	-0.6778	-0.5841	-0.0937	0.4667	0.8642	0.9821	(-2.5483, 1.3802)
13	0.7719	0.8665	-0.0946	0.4775	0.6944	0.8427	(-0.8190, 2.5520)
14	3.5018	3.806	-0.3042	0.3821	0.4592	0.5974	(2.6112, 5.0008)
15	4.214	3.6797	0.5343	0.4154	0.8036	0.9046	(1.8705, 5.4889)
16	3.9699	2.4846	1.4853	0.4893	0.7404	0.8874	(0.7097, 4.2594)

Table 24. Individual Effect of Errors on Bias, Standard Deviation, and Root Mean Square Error of Undercount Rate for the U.S. When all Other Errors Are Assumed to Be Zero
Estimated Undercount Rate = 1.18

Error Source	Bias	Std. Dev.	RMSE
E-sample collection error	-0.22	0.04	0.22
E-sample processing error	0.22	0.03	0.22
P-sample matching error	0.25	0.03	0.25
P-sample collection error	0.43	0.05	0.43
P-sample fabrication error	0.03	0.02	0.03
Sampling error	0.00	0.13	0.13
Correlation bias	-0.39	0.00	0.39
Ratio Estimator bias	0.01	0.00	0.01
Imputation	0.00	0.02	0.02

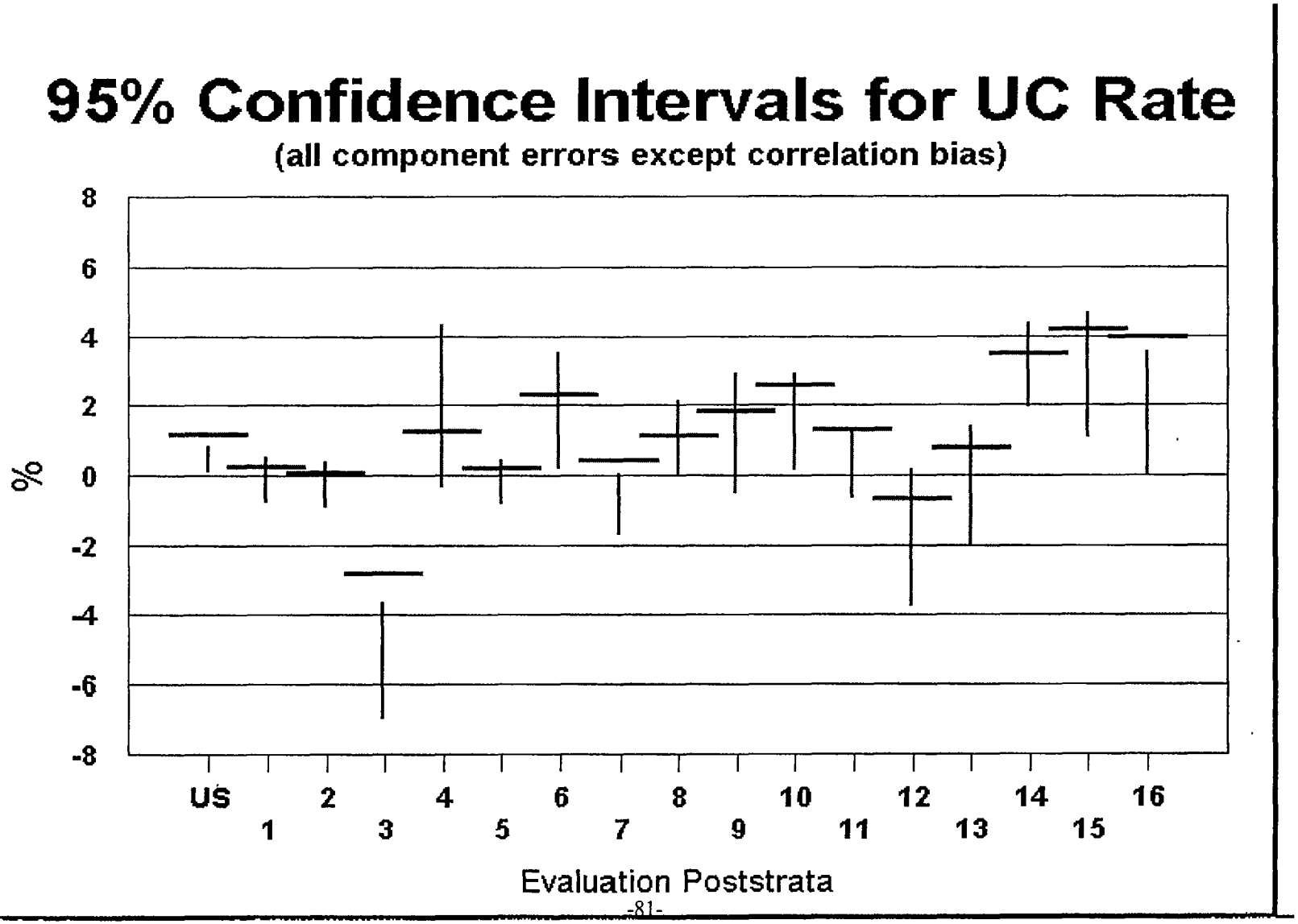

Figure 1. 95% confidence interval for net undercount rate assuming no correlation bias.

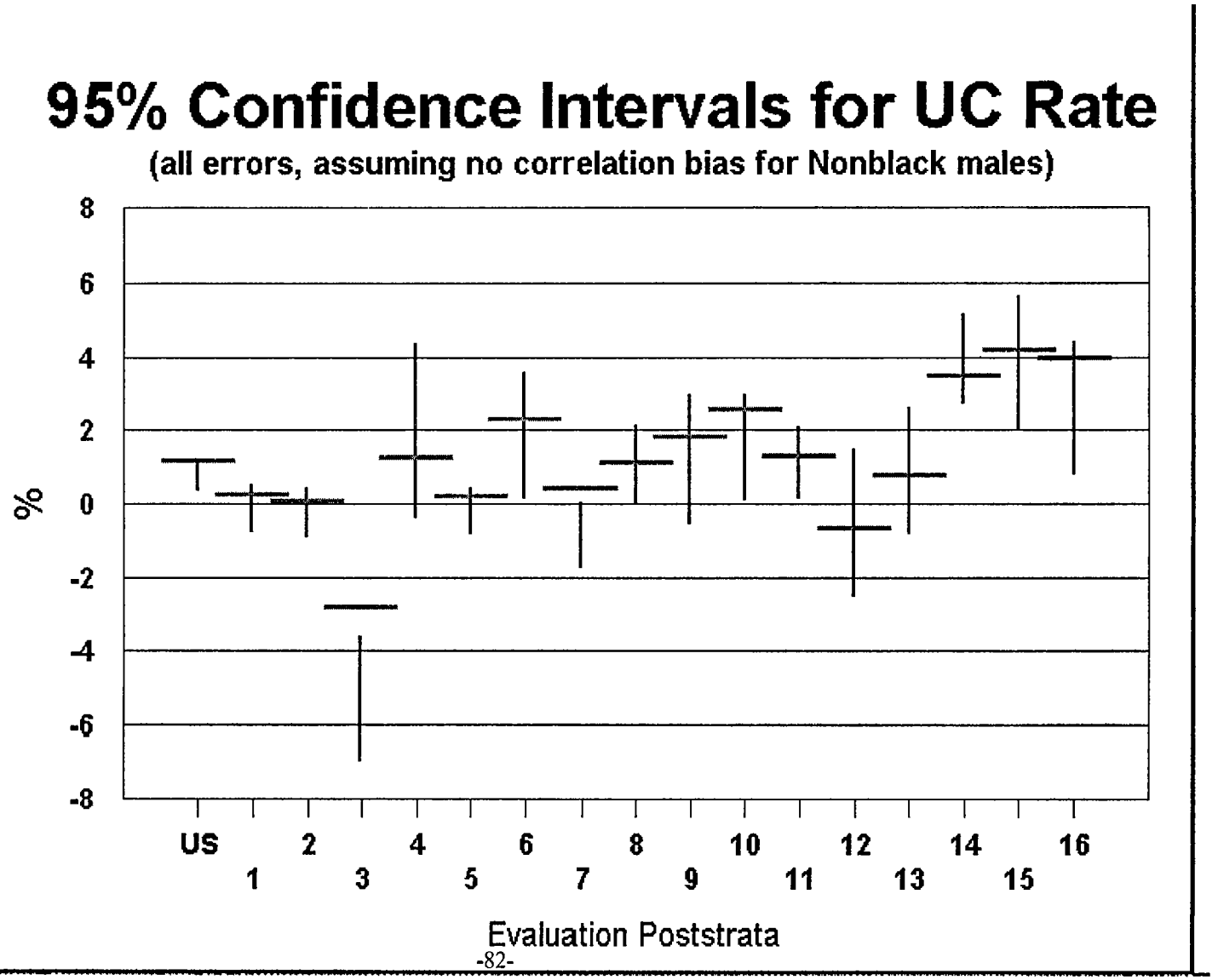

Figure 2. 95% confidence interval for net undercount rate assuming no correlation bias for Nonblack males.

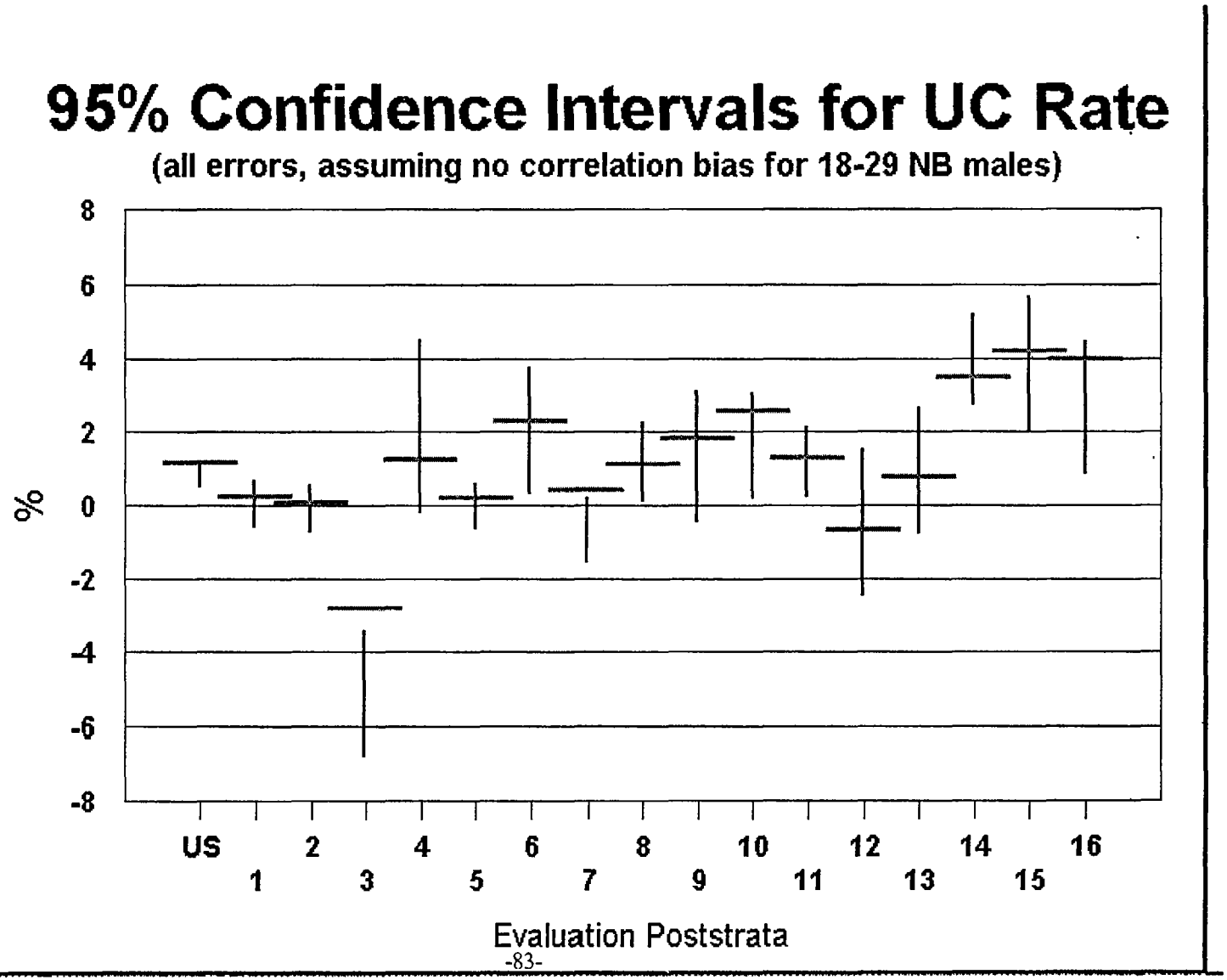

Figure 3. 95% confidence interval for net undercount rate assuming no correlation bias for 18-29 Nonblack males.

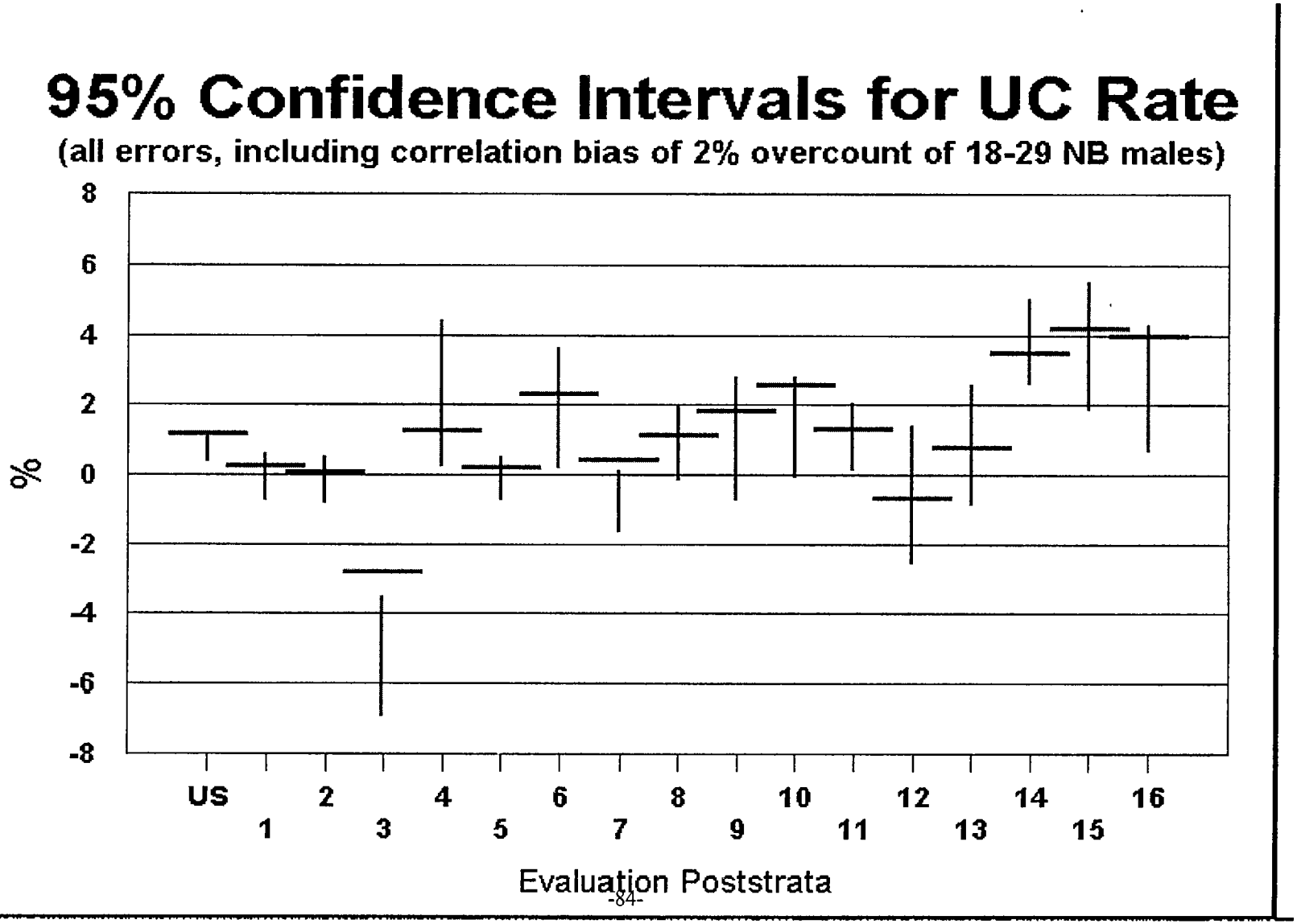

Figure 4. 95% confidence interval for net undercount rate including correlation bias of 2% overcount for 18-29 Nonblack males.

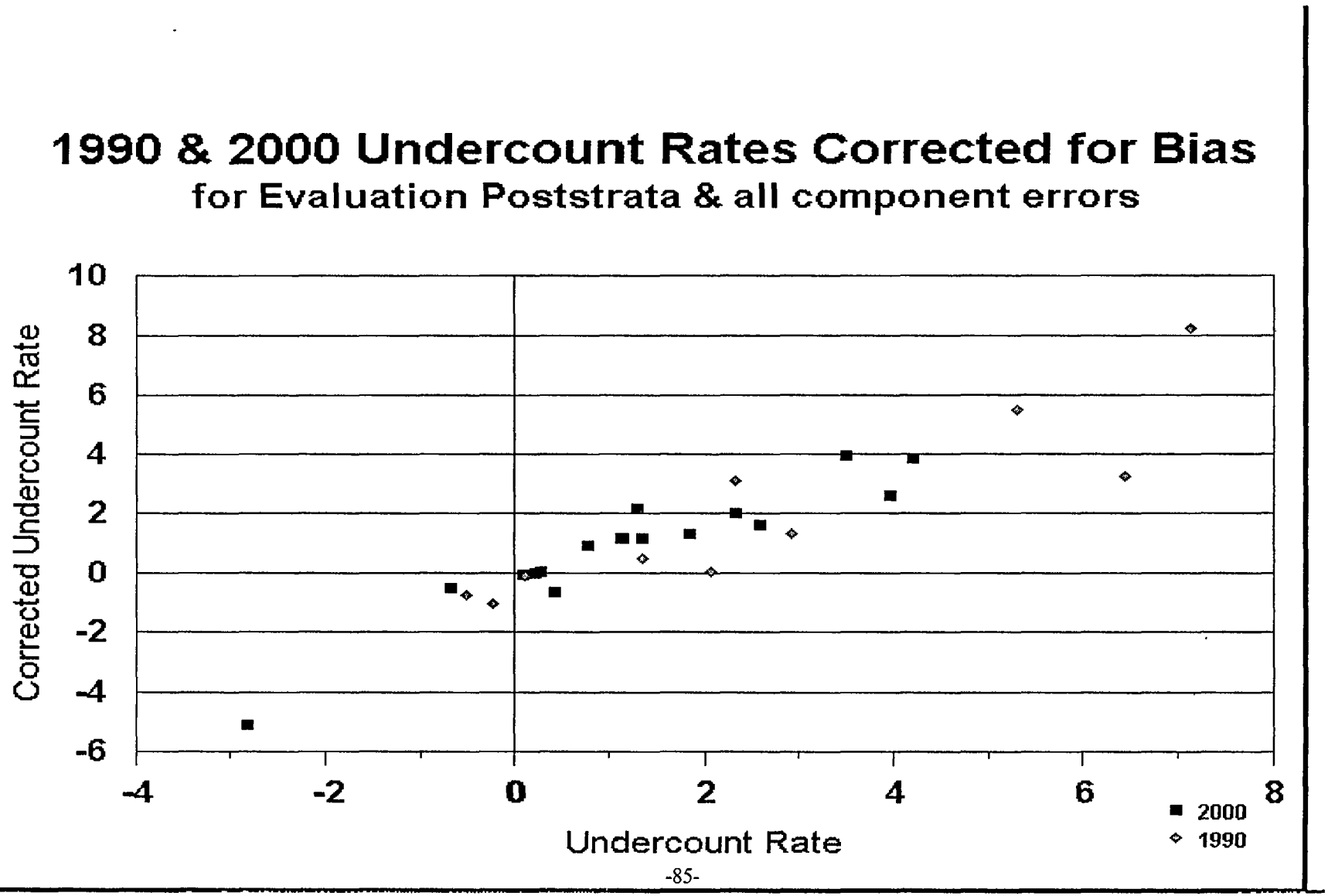

Figure 5a. 1990 & 2000 Undercount Rates Corrected for Bias for Evaluation Poststrata and All Component Errors.

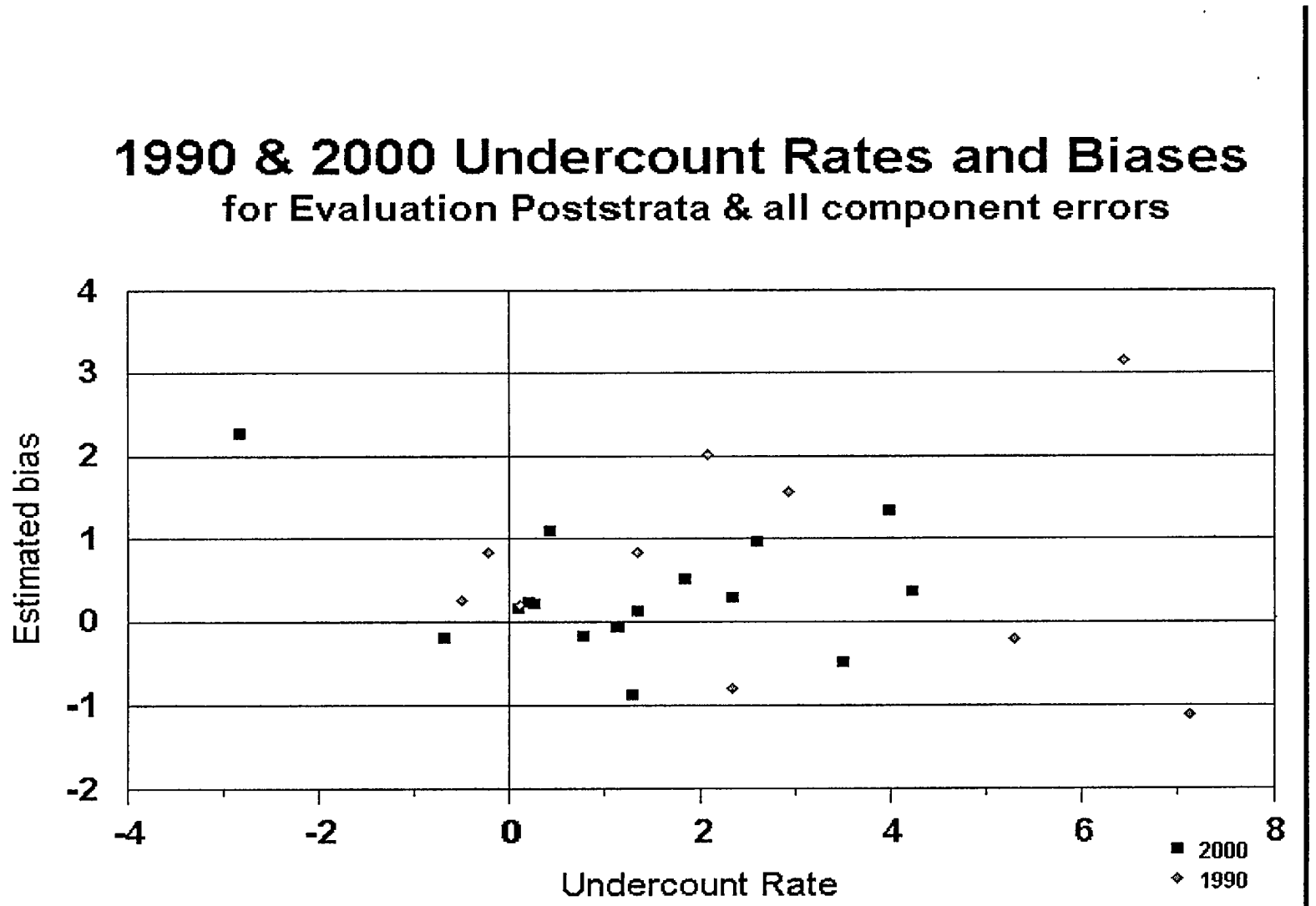

Figure 5b. 1990 & 2000 Undercount Rates and Biases for Evaluation Poststrata and All Component Errors.